METHYL-D-ERYTHRITOL PHOSPHATE PATHWAY GENES

This application claims the benefit under 35 U.S.C. § 119(e) of U.S. Provisional Patent Application Serial No. 60/223,483 filed August 7, 2000, which application is herein incorporated by reference.

A paper copy of the Sequence Listing and a computer readable form of the sequence listing on diskette, containing the file named 16516-107 seq listing.txt, which is 133,010 bytes in size (measured in MS-DOS), and which was created on August 6, 2001, are herein incorporated by reference.

The present invention is in the field of plant genetics and biochemistry. More specifically, the invention relates to genes associated with the methyl-D-erythritol phosphate (MEP) pathway. The present invention provides and includes nucleic acid molecules, proteins, and antibodies associated with the genes of the MEP pathway and also provides methods utilizing such agents, for example in gene isolation, gene analysis and the production of transgenic plants. Moreover, the present invention includes transgenic plants modified to express proteins associated with the MEP pathway and methods for the production of products from the MEP pathway.

Tocopherols are an important component of mammalian diets. Epidemiological evidence indicates that tocopherol supplementation can result in decreased risk for cardiovascular disease and cancer, can aid in immune function, and is associated with prevention or retardation of a number of degenerative disease processes in humans. Tocopherols function, in part, by stabilizing the lipid bilayer of biological membranes, reducing polyunsaturated fatty acid (PUFA) free radicals generated by lipid oxidation, and scavenging oxygen free radicals, lipid peroxy radicals and singlet oxygen species.

α-Tocopherol, often referred to as vitamin E, belongs to a class of lipid-soluble antioxidants that includes α, β, γ, and δ-tocopherols and α, β, γ, and δ-tocotrienols. Although α, β, γ, and δ-tocopherols and α, β, γ, and δ-tocotrienols are sometimes referred to collectively as "vitamin E", vitamin E is more appropriately defined chemically as α-tocopherol. α-Tocopherol is significant for human health, in part because it is readily absorbed and retained by the body, and therefore has a higher degree of bioactivity than other tocopherol species. However, other tocopherols such as β, γ, and δ-tocopherols, also have significant health and nutritional benefits.

Tocopherols are primarily synthesized only by plants and certain other photosynthetic organisms, including cyanobacteria. As a result, mammalian dietary tocopherols are obtained almost exclusively from these sources. Plant tissues vary considerably in total tocopherol content and tocopherol composition, with α-tocopherol the predominant tocopherol species found in green, photosynthetic plant tissues. Leaf tissue can contain from 10-50 μg of total tocopherols per gram fresh weight, but most of the world's major staple crops (*e.g.*, rice, corn, wheat, potato) produce low to extremely low levels of total tocopherols, of which only a small percentage is α-tocopherol. Oil seed crops generally contain much higher levels of total tocopherols, but α-tocopherol is present only as a minor component in most oilseeds.

The recommended human daily dietary intake of 15-30 mg of vitamin E is quite difficult to achieve from the average American diet. For example, it would take over 750 grams of spinach leaves in which α-tocopherol comprises 60% of total tocopherols, or 200-400 grams of soybean oil to satisfy this recommended daily vitamin E intake. While it is possible to augment the diet with supplements, most of these supplements contain primarily synthetic vitamin E, having eight stereoisomers, whereas natural vitamin E is predominantly composed of only a single isomer. Furthermore, supplements tend to be relatively expensive, and the general population is disinclined to take vitamin supplements on a regular basis.

In addition to the health benefits of tocopherols, increased α-tocopherol levels in crops have been associated with enhanced stability and extended shelf life of fresh and processed plant products. Further, tocopherol supplementation of swine, beef, and poultry feeds has been shown to significantly increase meat quality and extend the shelf life of post-processed meat products by retarding post-processing lipid oxidation, which contributes to undesirable flavor components.

Tocopherols are a member of the class of compounds referred to as the isoprenoids. Other isoprenoids include carotenoids, gibberellins, terpenes, chlorophyll and abscisic acid. The chloroplasts of higher plants exhibit interconnected biochemical pathways leading to secondary metabolites including tocopherols. One tocopherol biosynthetic pathway in higher plants involves condensation of homogentisic acid and phytylpyrophosphate to form 2-methyl-6 phytylplastoquinol.

This plant tocopherol pathway can be divided into four parts: 1) synthesis of homogentisic acid, which contributes to the aromatic ring of tocopherol; 2) synthesis of phytylpyrophosphate, which contributes to the side chain of tocopherol; 3) joining of HGA and phytylpyrophosphate via a prenyltransferase followed by a subsequent cyclization; 4) and S-adenosyl methionine-dependent methylation of an aromatic ring, which affects the relative abundance of each of the tocopherol species.

Homogentisic acid (HGA) is the common precursor to both tocopherols and plastoquinones. In at least some bacteria the synthesis of HGA is reported to occur via the conversion of chorismate to prephenate and then to p-hydroxyphenylpyruvate via a bifunctional prephenate dehydrogenase. Examples of bifunctional bacterial prephenate dehydrogenase enzymes include the proteins encoded by the *tyrA* genes of *Erwinia herbicola* and *Escherichia coli.* The *tyrA* gene product catalyzes the production of prephenate from chorismate, as well as the subsequent dehydrogenation of prephenate to form p-hydroxyphenylpyruvate (p-HPP), the immediate precursor to HGA. p-HPP is then converted to HGA by hydroxyphenylpyruvate dioxygenase (HPPD). In contrast,

plants are believed to lack prephenate dehydrogenase activity, and it is generally believed that the synthesis in plants of HGA from chorismate occurs via the synthesis and conversion of the intermediate arogenate. Because pathways involved in HGA synthesis are also responsible for tyrosine formation, any alterations in these pathways can also result in the alteration in tyrosine synthesis and the synthesis of other aromatic amino acids.

HGA is then combined with either phytyl-pyrophosphate or solanyl-pyrophosphate by phytyl/prenyl transferase to form methyl-plastoquinols, which are precursors to plastoquinones and tocopherols. The major structural difference between each of the tocopherol species is the position of the methyl groups around the phenyl ring. This methylation process is S-adenosyl methionine-dependent. Methyl Transferase 1 (MT1) catalyzes the formation of plastoquinol-9 and -tocopherol by methylation of the 7 position. Subsequent methylation at the 5 position of -tocopherol by -tocopherol methyl-transferase generates the biologically active -tocopherol.

Phytylpyrophosphate, which is the central constituent of the tocopherol side chain, is formed from geranylgeranyldiphosphate (GGDP). GGDP is itself produced via a biosynthetic pathway in which isopentenyl diphosphate (IPP) plays a major role. IPP is a central intermediate in the production of isoprenoids. Two pathways that generate IPP have been reported: a cytoplasmic-based pathway referred to as the mevalonate pathway; and a plastid-based pathway referred to as the MEP pathway. The cytoplasmic-based pathway involves the enzymes acetoacetyl CoA thiolase, HMGCoA synthase, HMGCoA reductase, mevalonate kinase, phosphomevalonate kinase, and mevalonate pyrophosphate decarboxylase.

Evidence for the existence of an alternative, plastid-based, isoprenoid biosynthetic pathway recently emerged from studies in the research groups of Rohmer and Arigoni, who found that the isotope labeling patterns observed in studies on certain eubacterial and plant terpenoids could not be explained in terms of the mevalonate pathway. Eisenreich

et al., *Chem. Bio.* 5:R221-233 (1998); Rohmer, *Prog. Drug. Res.* 50:135-154 (1998); Rohmer, 2 *Comprehensive Natural Products Chemistry* 45-68, Barton and Nakanishi (eds.), Pergamon Press, Oxford, England (1999). Arigoni and coworkers subsequently showed that 1-deoxyxylulose, or a derivative thereof, serves as an intermediate of the novel pathway, now referred to as the MEP pathway. Rohmer *et al.*, *Biochem. J.* 295:517-524 (1993); Schwarz, Ph.D. thesis, Eidgenössiche Technische Hochschule, Zurich, Switzerland (1994).

In the first step of the MEP pathway, DXP synthase, an enzyme encoded by the *dxs* gene, catalyzes the formation of 1-deoxy-D-xylulose-5-phosphate (DXP) from one molecule each of D-glyceraldehyde-3-phosphate and pyruvate. DXP is then converted into 2-*C*-methyl-D-erythritol-4-phosphate (MEP) by DXP reductoisomerase, which is encoded by the *dxr* gene. The conversion of MEP into 4-diphosphocytidyl-2-C-methyl-D-erythritol (CDP-ME) is catalyzed by CDP-ME synthase, which is encoded by the *ygbP* gene. CDP-ME kinase, which is encoded by the *ychB* gene, catalyzes the conversion of CDP-ME into 4-diphosphocytidyl-2-C-methyl-D-erythritol 2-phosphate (CDP-MEP). CDP-MEP is then converted into 2-C-methyl-D-erythritol-2,4-cyclodiphosphate by ME-CDP synthase, which is encoded by the *ygbB* gene. The *ygbP* and *ygbB* genes are tightly linked on the *E. coli* genome. Herz *et al.*, *PNAS* 97(6):2485-2490 (2000).

Identification of further genes included in the MEP pathway will provide new approaches to increasing tocopherol levels in plants, which is a topic of the present application.

SUMMARY OF THE INVENTION

The present invention provides a novel gene essential to the MEP pathway: *gcpE*. *gcpE* is tightly linked to *ygbP* and *ygbB*. Expression of GCPE (protein) in organisms such as plants can increase the levels of tocopherol substrates such as isopentyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) biosynthesis. The present

invention also provides transgenic organisms expressing a GCPE protein, which can nutritionally enhance food and feed sources.

In particular, the present invention includes and provides a substantially purified nucleic acid molecule that encodes a protein comprising an amino acid sequence selected from the group consisting of SEQ ID NOs: 4 and 48 through 50. The present invention also includes and provides a substantially purified nucleic acid molecule that encodes a protein comprising an amino acid sequence of SEQ ID NO: 4. Further provided by the present invention is a substantially purified nucleic acid molecule that encodes a protein comprising an amino acid sequence of SEQ ID NO: 48.

The present invention includes and provides a substantially purified nucleic acid molecule that encodes a protein comprising an amino acid sequence of SEQ ID NO: 49. The present invention also includes and provides a substantially purified nucleic acid molecule that encodes a protein comprising an amino acid sequence of SEQ ID NO: 50. Further provided by the present invention is a substantially purified nucleic acid molecule that encodes a GCPE protein, where the nucleic acid molecule comprises a nucleic acid sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof.

The present invention includes and provides a recombinant nucleic acid molecule comprising as operably linked components: (A) a promoter; and (B) a heterologous nucleic acid molecule that encodes an amino sequence selected from the group consisting of SEQ ID NOs: 4 and 48 through 50. The present invention also includes and provides transformed cells comprising such nucleic acid molecules.

Further provided by the present invention is a transgenic plant comprising a recombinant nucleic acid molecule comprising as operably linked components: (A) a promoter; and (B) a heterologous nucleic acid molecule that encodes an amino sequence selected from the group consisting of SEQ ID NOs: 4 and 48 through 50.

The present invention includes and provides such a transgenic plant that exhibits an increased tocopherol level relative to a plant with a similar genetic background but lacking the recombinant nucleic acid molecule. Also provided are seeds derived from such transgenic plants, and oil derived from such seeds. The present invention includes and provides such a transgenic plant that exhibits an increased monoterpene level relative to a plant with a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a transgenic plant that exhibits an increased carotenoid level relative to a plant with a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a transgenic plant that exhibits an increased tocotrienol level relative to a plant with a similar genetic background but lacking the recombinant nucleic acid molecule.

The present invention includes and provides such a transgenic plant that produces a seed with an increased tocopherol level relative to a plant with a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a transgenic plant that produces a seed with an increased monoterpene level relative to a plant with a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a transgenic plant that produces a seed with an increased carotenoid level relative to a plant with a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a transgenic plant which produces a seed with an increased tocotrienol level relative to a plant with a similar genetic background but lacking the recombinant nucleic acid molecule.

The present invention includes and provides a recombinant nucleic acid molecule comprising as operably linked components: (A) an exogenous promoter; and (B) a nucleic acid sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5

through 47, and complements thereof. The present invention also includes and provides transformed cells comprising such nucleic acid molecules.

Further provided by the present invention is a transgenic plant comprising a recombinant nucleic acid molecule comprising as operably linked components: (A) an exogenous promoter; and (B) a nucleic acid sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof. The present invention includes and provides such a transgenic plant which is selected from the group consisting of *Brassica campestris, Brassica napus,* canola, castor bean, coconut, cotton, crambe, linseed, maize, mustard, oil palm, peanut, rapeseed, rice, safflower, sesame, soybean, sunflower, and wheat. The present invention includes and provides such a trangenic plant which is selected from the group consisting of coconut, crambe, maize, oil palm, peanut, rapeseed, safflower, sesame, soybean, and sunflower.

The present invention further includes and provides a seed derived from such a transgenic plant. Also provided are oil and meal derived from such seeds. The present invention includes and provides such a seed which exhibits an increased tocopherol level relative to seed from a plant having a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a seed which exhibits an increased -tocopherol level relative to seed from a plant having a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a seed which exhibits an increased monoterpene level relative to seed from a plant having a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a seed which exhibits an increased carotenoid level relative to seed from a plant having a similar genetic background but lacking the recombinant nucleic acid molecule. The present invention includes and provides such a seed which exhibits an increased tocotrienol level relative to seed from a plant having a similar genetic background but lacking the recombinant nucleic acid molecule.

The present invention includes and provides a recombinant nucleic acid molecule comprising as operably linked components: (A) a promoter that functions in a plant cell to cause production of an mRNA molecule; and (B) a nucleic acid sequence that hybridizes under moderate stringency conditions to a nucleic acid sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof.

The present invention includes and provides a recombinant nucleic acid molecule comprising as operably linked components: (A) a promoter that functions in a plant cell to cause production of an mRNA molecule; and (B) a nucleic acid sequence that has greater than 85% identity to a nucleic acid sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof.

The present invention includes and provides a substantially purified protein comprising an amino acid sequence selected from the group consisting of SEQ ID NOs: 4, 48, and 49. The present invention also includes and provides an antibody capable of specifically binding a protein comprising an amino acid sequence selected from the group consisting of SEQ ID NOs: 4, 48 and 49.

The present invention includes and provides a transgenic plant comprising a nucleic acid molecule that encodes a GCPE protein, where the nucleic acid molecule comprises a promoter operably linked to a heterologous nucleic acid sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof. The present invention includes and provides such a transgenic plant where the the promoter is a seed-specific promoter. The present invention includes and provides such a transgenic plant where the seed-specific promoter is selected from the group consisting of napin, phaseolin, zein, soybean trypsin inhibitor, ACP, stearoyl-ACP desaturase, soybean a' subunit of b-conglycinin (soy 7s), and oleosin promoters.

The present invention includes and provides such a transgenic plant, where the plant exhibits an increased isoprenoid compound level relative to a plant with a similar

genetic background but lacking the heterologous nucleic acid sequence. The present invention includes and provides such a transgenic plant, where the isoprenoid compound is selected from the group consisting of tocotrienols, tocopherols, terpenes, gibberellins, carotenoids, and xanthophylls. The present invention includes and provides such a transgenic plant, where the isoprenoid compound is a monoterpene. The present invention includes and provides such a transgenic plant, where the isoprenoid compound is selected from the group consisting of IPP and DMAPP. The present invention includes and provides such a transgenic plant, where the plant exhibits an increased tocopherol level relative to a plant with a similar genetic background but lacking the heterologous nucleic acid sequence. Also included and provided are feedstock, plant parts, and seeds derived from such plants. Further provided are containers of such seeds.

The present invention includes and provides a method of producing a transgenic plant with an increased isoprenoid compound level comprising: (A) transforming the plant with a nucleic acid molecule to produce a transgenic plant, where the nucleic acid molecule comprises a nucleic acid sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof; and (B) growing the transgenic plant.

The present invention includes and provides a method of producing a transgenic plant having seed with an increased isoprenoid compound level comprising: (A) transforming the plant with a nucleic acid molecule to produce a transgenic plant, where the nucleic acid molecule encodes a protein with an amino acid sequence selected from the group consisting of SEQ ID NOs: 4 and 48-50; and (B) growing the transgenic plant.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 sets forth chemical compounds that were determined as non-GCPE reaction products.

Figure 2 sets forth the diacetate of 2-methylbut-2-ene-1,4-diol.

Figure 3 sets forth (E)-1-(4-hydroxy-3-methylbut-2-enyl) diphosphate.

Figure 4 sets forth an alignment between proteins encoded by the *gcpE* gene from *E. coli* (SEQ ID NO: 78) and clone 135H1 from *A. thaliana* (SEQ ID NO: 79).

Figure 5 sets forth cloning of a truncated *Arabidopsis* cDNA to create pQE-AGH.

DESCRIPTION OF THE NUCLEIC AND AMINO ACID SEQUENCES

SEQ ID NO: 1 is an *Arabidopsis thaliana* nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 2 is a rice nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 3 is an *E. coli* nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 4 is an amino acid sequence derived from a rice *gcpE* gene.

SEQ ID NO: 5 is a partial *A. thaliana* nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 6 is a partial soybean nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 7 is a partial tomato nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 8 is a partial *Mesembryanthemun crystallinum* nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 9 is a partial rice nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 10 is a partial maize nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 11 is a partial Loblolly pine nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 12 is a partial *Physcomitrella patens* nucleotide sequence of a *gcpE* gene.

SEQ ID NOs: 13 through 20 are partial *A. thaliana* nucleotide sequences of a *gcpE* gene.

SEQ ID NOs: 21 through 32 are partial maize nucleotide sequences of a *gcpE* gene.

SEQ ID NOs: 33 through 46 are partial soybean nucleotide sequences of a *gcpE* gene.

SEQ ID NO: 47 is a partial *Brassica napus* nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 48 is an amino acid sequence derived from an *A. thaliana gcpE* gene.

SEQ ID NO: 49 is an amino acid sequence derived from a rice *gcpE* gene.

SEQ ID NO: 50 is an amino acid sequence derived from an *E. coli gcpE* gene.

SEQ ID NOs: 51 through 77 are primer nucleotide sequences.

SEQ ID NO: 78 is an *E. coli* amino acid sequence derived from the *gcpE* gene.

SEQ ID NO: 79 is an *A. thaliana* amino acid sequence derived from clone 135H1.

SEQ ID NO: 80 is a partial *A. thaliana* nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 81 is an amino acid sequence derived from an *A. thaliana gcpE* gene.

SEQ ID NO: 82 is a partial *A. thaliana* nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 83 is an amino acid sequence derived from an *A. thaliana gcpE* gene.

SEQ ID NO: 84 is a partial *A. thaliana* nucleotide sequence of a *gcpE* gene.

SEQ ID NO: 85 is an amino acid sequence derived from an *A. thaliana gcpE* gene.

DEFINITIONS

The following definitions are provided as an aid to understanding the detailed description of the present invention.

The abbreviation "EP" refers to patent applications and patents published by the European Patent Office, and the term "WO" refers to patent applications published by the World Intellectual Property Organization. "PNAS" refers to *Proc. Natl. Acad. Sci. (U.S.A.)*.

"Amino acid" and "amino acids" refer to all naturally occurring L-amino acids. This definition is meant to include norleucine, norvaline, ornithine, homocysteine, and homoserine.

"Chromosome walking" means a process of extending a genetic map by successive hybridization steps.

The phrases "coding sequence," "structural sequence," and "structural nucleic acid sequence" refer to a physical structure comprising an orderly arrangement of nucleic acids. The coding sequence, structural sequence, and structural nucleic acid sequence may be contained within a larger nucleic acid molecule, vector, or the like. In addition, the orderly arrangement of nucleic acids in these sequences may be depicted in the form of a sequence listing, figure, table, electronic medium, or the like.

A nucleic acid molecule is said to be the "complement" of another nucleic acid molecule if they exhibit complete complementarity, *i.e.*, every nucleotide of one of the molecules is complementary to a nucleotide of the other. Two molecules are "minimally complementary" if they can hybridize to one another with sufficient stability to remain annealed to one another under at least conventional "low-stringency" conditions. Similarly, the molecules are "complementary" if they can hybridize to one another with sufficient stability to remain annealed to one another under conventional "high-stringency" conditions. Conventional stringency conditions are described by Sambrook *et al., Molecular Cloning: A Laboratory Manual*, Second Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (1989); Haymes *et al., Nucleic Acid Hybridization, A Practical Approach*, IRL Press, Washington, DC (1985).

The phrases "DNA sequence," "nucleic acid sequence," and "nucleic acid molecule" refer to a physical structure comprising an orderly arrangement of nucleic acids. The DNA sequence or nucleic acid sequence may be contained within a larger nucleic acid molecule, vector, or the like. In addition, the orderly arrangement of nucleic acids in these sequences may be depicted in the form of a sequence listing, figure, table,

electronic medium, or the like. "Nucleic acid" refers to deoxyribonucleic acid (DNA) and ribonucleic acid (RNA).

An "elite soybean line" is any soybean line that has resulted from breeding and selection for superior agronomic performance. Elite soybean lines are commercially available to farmers or soybean breeders, *e.g.*, HARTZ™ variety H4452 Roundup Ready™ (HARTZ SEED, Stuttgart, Arkansas, USA); QP4544 (Asgrow Seeds, Des Moines, Iowa, USA); DeKalb variety CX445 (DeKalb, Illinois).

"Exogenous genetic material" is any genetic material, whether naturally occurring or otherwise, from any source that is capable of being inserted into any organism.

The term "expression" refers to the transcription of a gene to produce the corresponding mRNA and translation of this mRNA to produce the corresponding gene product (*i.e.*, a peptide, polypeptide, or protein). The term "expression of antisense RNA" refers to the transcription of a DNA to produce a first RNA molecule capable of hybridizing to a second RNA molecule. Formation of the RNA-RNA hybrid inhibits translation of the second RNA molecule to produce a gene product.

"Fungi" as used herein includes the phyla *Ascomycota, Basidiomycota, Chytridiomycota* and *Zygomycota*, as well as the *Oomycota* and all mitosporic fungi, and "filamentous fungi" include all filamentous forms of the subdivision *Eumycota* and *Oomycota*. These terms are defined in Hawksworth *et al.*, in: Ainsworth and Bisby's *Dictionary of The Fungi*, 8th edition, CAB International, University Press, Cambridge, UK (1995).

"Homology" refers to the level of similarity between two or more nucleic acid or amino acid sequences in terms of percent of positional identity (*i.e.*, sequence similarity or identity). Homology also refers to the concept of similar functional properties among different nucleic acids or proteins.

As used herein, a "homolog protein" molecule or fragment thereof is a counterpart protein molecule or fragment thereof in a second species (*e.g.*, maize GCPE is a homolog

of *Arabidopsis* GCPE). A homolog can also be generated by molecular evolution or DNA shuffling techniques, so that the molecule retains at least one functional or structure characteristic of the original protein (*see, e.g.*, U.S. Patent No. 5,811,238).

The phrase "heterologous" refers to the relationship between two or more nucleic acid or protein sequences that are derived from different sources. For example, a promoter is heterologous with respect to a coding sequence if such a combination is not normally found in nature. In addition, a particular sequence may be "heterologous" with respect to a cell or organism into which it is inserted (*i.e.* does not naturally occur in that particular cell or organism).

"Hybridization" refers to the ability of a strand of nucleic acid to join with a complementary strand via base pairing. Hybridization occurs when complementary nucleic acid sequences in the two nucleic acid strands contact one another under appropriate conditions.

The "MEP pathway" is the pathway associated with the biosynthesis of isopentenyl diphosphate or dimethylallyldiphosphate where deoxy-D-xylulose-5-phosphate or a derivative thereof serves as an intermediate.

The phrase "operably linked" refers to the functional spatial arrangement of two or more nucleic acid regions or nucleic acid sequences. For example, a promoter region may be positioned relative to a nucleic acid sequence such that transcription of a nucleic acid sequence is directed by the promoter region. Thus, a promoter region is "operably linked" to the nucleic acid sequence.

"Phenotype" refers to traits exhibited by an organism resulting from the interaction of genotype and environment, such as disease resistance, pest tolerance, environmental tolerance such as tolerance to abiotic stress, male sterility, quality improvement or yield *etc.*

"Polyadenylation signal" or "polyA signal" refers to a nucleic acid sequence located 3' to a coding region that promotes the addition of adenylate nucleotides to the 3' end of the mRNA transcribed from the coding region.

The term "promoter" or "promoter region" refers to a nucleic acid sequence, usually found upstream (5') to a coding sequence, which is capable of directing transcription of a nucleic acid sequence into mRNA. The promoter or promoter region typically provide a recognition site for RNA polymerase and the other factors necessary for proper initiation of transcription. As contemplated herein, a promoter or promoter region includes variations of promoters derived by inserting or deleting regulatory regions, subjecting the promoter to random or site-directed mutagenesis, *etc.* The activity or strength of a promoter may be measured in terms of the amounts of RNA it produces, or the amount of protein accumulation in a cell or tissue, relative to a promoter whose transcriptional activity has been previously assessed.

The term "protein" or "peptide molecule" includes any molecule that comprises five or more amino acids. It is well known in the art that proteins may undergo modification, including post-translational modifications, such as, but not limited to, disulfide bond formation, glycosylation, phosphorylation, or oligomerization. Thus, as used herein, the term "protein" or "peptide molecule" includes any protein that is modified by any biological or non-biological process.

A "protein fragment" is a peptide or polypeptide molecule whose amino acid sequence comprises a subset of the amino acid sequence of that protein. A protein or fragment thereof that comprises one or more additional peptide regions not derived from that protein is a "fusion" protein.

"Recombinant vector" refers to any agent such as a plasmid, cosmid, virus, autonomously replicating sequence, phage, or linear single-stranded, circular single-stranded, linear double-stranded, or circular double-stranded DNA or RNA nucleotide

sequence. The recombinant vector may be derived from any source and is capable of genomic integration or autonomous replication.

"Regeneration" refers to the process of growing a plant from a plant cell or plant tissue (e.g., plant protoplast or explant).

"Regulatory sequence" refers to a nucleotide sequence located upstream (5'), within, or downstream (3') to a coding sequence. Transcription and expression of the coding sequence is typically impacted by the presence or absence of the regulatory sequence.

An antibody or peptide is said to "specifically bind" to a protein or peptide molecule of the invention if such binding is not competitively inhibited by the presence of non-related molecules.

"Substantially homologous" refers to two sequences which are at least 90% identical in sequence, as measured by the BestFit program described herein (Version 10; Genetics Computer Group, Inc., University of Wisconsin Biotechnology Center, Madison, WI), using default parameters.

"Substantially purified" refers to a molecule separated from substantially all other molecules normally associated with it in its native state. More preferably a substantially purified molecule is the predominant species present in a preparation. A substantially purified molecule may be greater than 60% free, preferably 75% free, more preferably 90% free, and most preferably 95% free from the other molecules (exclusive of solvent) present in the natural mixture. The term "substantially purified" is not intended to encompass molecules present in their native state.

"Transcription" refers to the process of producing an RNA copy from a DNA template. "Transformation" refers to the introduction of nucleic acid into a recipient host. The term "host" refers to bacteria cells, fungi, animals or animal cells, plants or seeds, or any plant parts or tissues including plant cells, protoplasts, calli, roots, tubers, seeds, stems, leaves, seedlings, embryos, and pollen.

"Transgenic" refers to organisms into which exogenous nucleic acid sequences are integrated. "Transgenic plant" refers to a plant where an introduced nucleic acid is stably introduced into a genome of the plant, for example, the nuclear or plastid genomes.

"Vector" refers to a plasmid, cosmid, bacteriophage, or virus that carries exogenous DNA into a host organism.

"Yeast" as used herein includes *Ascosporogenous* yeast (*Endomycetales*), *Basidiosporogenous* yeast and yeast belonging to the *Fungi Imperfecti* (*Blastomycetes*), as defined in Skinner *et al.* (1980).

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

One skilled in the art may refer to general reference texts for detailed descriptions of known techniques discussed herein or equivalent techniques. These texts include Ausubel *et al., Current Protocols in Molecular Biology*, John Wiley and Sons, Inc. (1995); Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual* (2d ed.), Cold Spring Harbor Press, Cold Spring Harbor, New York (1989); Birren *et al.*, *Genome Analysis: A Laboratory Manual*, volumes 1 through 4, Cold Spring Harbor Press, Cold Spring Harbor, New York (1997-1999); *Plant Molecular Biology: A Laboratory Manual*, Clark (ed.), Springer, New York (1997); Richards et al., *Plant Breeding Systems* (2d ed.), Chapman & Hall, The University Press, Cambridge (1997); and Maliga *et al.*, *Methods in Plant Molecular Biology*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1995). These texts can, of course, also be referred to in making or using an aspect of the invention.

Utilizing a methodology for the isolation and characterization of essential MEP pathway genes, an essential and novel gene, termed *gcpE,* was isolated. *gcpE* is tightly linked to *ygbP* and *ygbB*, which are other MEP pathway genes. As an essential MEP pathway component, enhanced expression or overexpression of GCPE in a variety of

organisms such as plants can result in higher levels of tocopherol precursors such as IPP and DMAPP and ultimately in enhanced levels of tocopherols in such organisms. Moreover, the present invention provides a number of agents, for example, nucleic acid molecules encoding a GCPE protein, and provides uses of such agents.

The agents of the invention will preferably be "biologically active" with respect to either a structural attribute, such as the capacity of a nucleic acid to hybridize to another nucleic acid molecule, or the ability of a protein to be bound by an antibody (or to compete with another molecule for such binding). Alternatively, such an attribute may be catalytic and thus involve the capacity of the agent to mediate a chemical reaction or response. The agents will preferably be substantially purified. The agents of the invention may also be recombinant.

It is understood that any of the agents of the invention can be substantially purified and/or be biologically active and/or recombinant. It is also understood that the agents of the invention may be labeled with reagents that facilitate detection of the agent, *e.g.*, fluorescent labels, chemical labels, modified bases, and the like.

A. Nucleic Acid Molecules

Agents of the invention include nucleic acid molecules. In a preferred aspect of the present invention the nucleic acid molecule comprises a nucleic acid sequence which encodes a GCPE protein. In a preferred embodiment, the GCPE protein is derived from an organism having a MEP pathway. Examples of GCPE proteins are those proteins having an amino acid sequence selected from the group consisting of SEQ ID NO: 4, 48, 49, or 50.

In another preferred aspect of the present invention the nucleic acid molecule comprises a nucleic acid sequence that is selected from: (1) any of SEQ ID NOs: 1 through 3, 5 through 47, complements thereof, or fragments of these sequences; (2) the group consisting of SEQ ID NOs: 1, 2, complements thereof, and fragments of these

sequences; (3) the group consisting of SEQ ID NOs: 1, 2, 3, complements thereof and fragments of these sequences; (4) the group consisting of SEQ ID NOs: 1, 2, 13 through 47, complements thereof and fragments of these sequences; (5) the group consisting of SEQ ID NOs: 5 through 12, complements thereof and fragments of these sequences; or (6) the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, complements thereof and fragments of these sequences.

In a further aspect of the present invention the nucleic acid molecule comprises a nucleic acid sequence encoding an amino acid sequence selected from: (1) any of SEQ ID NOs: 4, 48, 49 or 50; (2) the group consisting of SEQ ID NO: 4, 48, and 49 and fragments of these sequences; or (3) the group consisting of SEQ ID NO: 4, 48, 49, 50 and fragments of these sequences.

It is understood that in a further aspect of the nucleic acid sequences of the present invention can encode a protein which differs from any of the proteins in that amino acid have been deleted, substituted or added without altering the function. For example, it is understood that codons capable of coding for such conservative amino acid substitutions are known in the art.

The present invention provides nucleic acid molecules that hybridize to the above-described nucleic acid molecules. Nucleic acid hybridization is a technique well known to those of skill in the art of DNA manipulation. The hybridization properties of a given pair of nucleic acids is an indication of their similarity or identity.

The nucleic acid molecules preferably hybridize, under low, moderate, or high stringency conditions, with a nucleic acid sequence selected from: (1) any of SEQ ID NOs: 1 through 3, 5 through 47, or complements thereof; (2) the group consisting of SEQ ID NOs: 1, 2, and complements thereof; (3) the group consisting of SEQ ID NOs: 1, 2, 3, and complements thereof; (4) the group consisting of SEQ ID NOs: 1, 2, 13 through 47, and complements thereof; (5) the group consisting of SEQ ID NOs: 5 through 12, and

complements thereof; or (6) the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof. Fragments of these sequences are also contemplated.

The hybridization conditions typically involve nucleic acid hybridization in about 0.1X to about 10X SSC (diluted from a 20X SSC stock solution containing 3 M sodium chloride and 0.3 M sodium citrate, pH 7.0 in distilled water), about 2.5X to about 5X Denhardt's solution (diluted from a 50X stock solution containing 1% (w/v) bovine serum albumin, 1% (w/v) ficoll, and 1% (w/v) polyvinylpyrrolidone in distilled water), about 10 mg/mL to about 100 mg/mL fish sperm DNA, and about 0.02% (w/v) to about 0.1% (w/v) SDS, with an incubation at about 20°C to about 70°C for several hours to overnight. The stringency conditions are preferably provided by 6X SSC, 5X Denhardt's solution, 100 mg/mL fish sperm DNA, and 0.1% (w/v) SDS, with an incubation at 55°C for several hours.

The hybridization is generally followed by several wash steps. The wash compositions generally comprise 0.1X to about 10X SSC, and 0.01% (w/v) to about 0.5% (w/v) SDS with a 15 minute incubation at about 20°C to about 70°C. Preferably, the nucleic acid segments remain hybridized after washing at least one time in 0.1X SSC at 65°C. For example, the salt concentration in the wash step can be selected from a low stringency of about 2.0 X SSC at 50°C to a high stringency of about 0.2 X SSC at 65°C. In addition, the temperature in the wash step can be increased from low stringency conditions at room temperature, about 22°C, to high stringency conditions at about 65°C. Both temperature and salt may be varied, or either the temperature or the salt concentration may be held constant while the other variable is changed.

Low stringency conditions may be used to select nucleic acid sequences with lower sequence identities to a target nucleic acid sequence. One may wish to employ conditions such as about 6.0 X SSC to about 10 X SSC, at temperatures ranging from about 20°C to about 55°C, and preferably a nucleic acid molecule will hybridize to one or more of the above-described nucleic acid molecules under low stringency conditions of

about 6.0 X SSC and about 45°C. In a preferred embodiment, a nucleic acid molecule will hybridize to one or more of the above-described nucleic acid molecules under moderately stringent conditions, for example at about 2.0 X SSC and about 65°C. In a particularly preferred embodiment, a nucleic acid molecule of the present invention will hybridize to one or more of the above-described nucleic acid molecules under high stringency conditions such as 0.2 X SSC and about 65°C.

In an alternative embodiment, the nucleic acid molecule comprises a nucleic acid sequence that is greater than 85% identical, and more preferably greater than 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, or 99% identical to a nucleic acid sequence selected from the group consisting of SEQ ID NO: 1 through 3 and 5 through 47, complements thereof, and fragments of any of these sequences.

The percent identity is preferably determined using the "Best Fit" or "Gap" program of the Sequence Analysis Software Package™ (Version 10; Genetics Computer Group, Inc., University of Wisconsin Biotechnology Center, Madison, WI). "Gap" utilizes the algorithm of Needleman and Wunsch to find the alignment of two sequences that maximizes the number of matches and minimizes the number of gaps. "BestFit" performs an optimal alignment of the best segment of similarity between two sequences and inserts gaps to maximize the number of matches using the local homology algorithm of Smith and Waterman. The percent identity calculations may also be performed using the Megalign program of the LASERGENE bioinformatics computing suite (default parameters, DNASTAR Inc., Madison, Wisconsin). The percent identity is most preferably determined using the "Best Fit" program using default parameters.

The present invention also provides nucleic acid molecule fragments that hybridize to the above-described nucleic acid molecules and complements thereof, fragments of nucleic acid molecules that exhibit greater than 80%, 85%, 90%, 95% or 99% sequence identity with the above-described nucleic acid molecules and complements thereof, or fragments of any of these molecules.

Fragment nucleic acid molecules may consist of significant portion(s) of, or indeed most of, the nucleic acid molecules of the invention. In an embodiment, the fragments are between about 3000 and about 1000 consecutive nucleotides, about 1800 and about 150 consecutive nucleotides, about 1500 and about 500 consecutive nucleotides, about 1300 and about 250 consecutive nucleotides, about 1000 and about 200 consecutive nucleotides, about 800 and about 150 consecutive nucleotides, about 500 and about 100 consecutive nucleotides, about 300 and about 75 consecutive nucleotides, about 100 and about 50 consecutive nucleotides, about 50 and about 25 consecutive nucleotides, or about 20 and about 10 consecutive nucleotides long of a nucleic molecule of the present invention.

In another embodiment, the fragment comprises at least 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 500, or 750 consecutive nucleotides of a nucleic acid sequence of the present invention.

Exemplary Uses

Nucleic acid molecules of the invention and fragments thereof may be employed to obtain other nucleic acid molecules from the same species (*e.g.*, nucleic acid molecules from maize may be utilized to obtain other nucleic acid molecules from maize). Exemplary nucleic acid molecules that may be obtained include, but are not limited to, nucleic acid molecules that encode the complete coding sequence of a protein and promoters and flanking sequences of such molecules, and nucleic acid molecules that encode for other isozymes or gene family members.

Nucleic acid molecules of the invention and fragments thereof may also be employed to obtain nucleic acid homologs. Such homologs include the nucleic acid molecules of other plants or other organisms, including the nucleic acid molecules that encode, in whole or in part, protein homologs of other plant species or other organisms,

or sequences of genetic elements, such as promoters and transcriptional regulatory elements.

Promoters that may be isolated include, but are not limited to promoters of cell enhanced, cell specific, tissue enhanced, tissue specific, developmentally or environmentally regulated expression profiles. Promoters obtained utilizing the nucleic acid molecules of the invention could also be modified to affect their control characteristics. Examples of such modifications would include but are not limited to enhancer sequences. Such genetic elements could be used to enhance gene expression of new and existing traits for crop improvement.

The above-described molecules can be readily obtained by using the above-described nucleic acid molecules or fragments thereof to screen cDNA or genomic libraries obtained from such plant species. These methods are known to those of skill in the art, as are methods for forming such libraries. In one embodiment, such sequences are obtained by incubating nucleic acid molecules of the present invention with members of genomic libraries and recovering clones that hybridize to such nucleic acid molecules thereof. In a second embodiment, methods of chromosome walking or inverse PCR may be used to obtain such sequences.

Any of a variety of methods may be used to obtain one or more of the above-described nucleic acid molecules. Automated nucleic acid synthesizers may be employed for this purpose. In lieu of such synthesis, the disclosed nucleic acid molecules may be used to define a pair of primers that can be used with the polymerase chain reaction to amplify and obtain any desired nucleic acid molecule or fragment.

In a preferred embodiment, nucleic acid molecules having SEQ ID NOs: 1 through 3 and 5 through 47, and complements thereof, and fragments of any of these sequences can be utilized to obtain such homologs. Such homolog molecules may differ in their nucleotide sequences from those found in one or more of SEQ ID NOs: 1 through 3, and 5 through 47 or complements thereof because complete complementarity is not

needed for stable hybridization. The nucleic acid molecules of the invention therefore also include molecules that, although capable of specifically hybridizing with the nucleic acid molecules may lack "complete complementarity."

In a preferred embodiment, the molecules are obtained from alfalfa, apple, *Arabidopsis,* banana, barley, *Brassica, Brassica campestris, Brassica napus*, broccoli, cabbage, canola, castor bean, chrysanthemum, citrus, coconut, coffee, cotton, crambe, cranberry, cucumber, Cuphea, dendrobium, dioscorea, eucalyptus, fescue, fir, garlic, gladiolus, grape, hordeum, lentils, lettuce, liliacea, linseed, maize, millet, muskmelon, mustard, oat, oil palm, oilseed rape, onion, an ornamental plant, papaya, pea, peanut, pepper, perennial ryegrass, *Phaseolus,* pine, poplar, potato, rapeseed (including Canola and High Erucic Acid varieties), rice, rye, safflower, sesame, sorghum, soybean, strawberry, sugarbeet, sugarcane, sunflower, tea, tomato, triticale, turf grasses, and wheat.

In a more preferred embodiment, the molecules are obtained from *Brassica campestris, Brassica napus,* canola, castor bean, coconut, cotton, crambe, linseed, maize, mustard, oil palm, peanut, rapeseed (including Canola and High Erucic Acid varieties), rice, safflower, sesame, soybean, sunflower, and wheat, and in a particularly preferred embodiment from coconut, crambe, maize, oil palm, peanut, rapeseed (including Canola and High Erucic Acid varieties), safflower, sesame, soybean, and sunflower.

The Sequence Analysis Software Package™ (Version 10; Genetics Computer Group, Inc., University of Wisconsin Biotechnology Center, Madison, WI) contains a number of other useful sequence analysis tools for identifying homologs of the presently disclosed nucleotide and amino acid sequences. For example, programs such as "BLAST", "FastA", "TfastA", "FastX", and "TfastX" can be used to search for sequences similar to a query sequence. *See, e.g.*, Altschul *et al.*, *Journal of Molecular Biology* 215: 403-410 (1990); Lipman and Pearson, *Science 227*:1435-1441 (1985); Pearson and Lipman, *85*:2444-2448 (1988); Pearson, "Rapid and Sensitive Sequence Comparison

with FASTP and FASTA" in *Methods in Enzymology* , (R. Doolittle, ed.), 183:63-98, Academic Press, San Diego, California, USA (1990).

Short nucleic acid sequences having the ability to specifically hybridize to complementary nucleic acid sequences may be produced and utilized in the present invention, *e.g.*, as probes to identify the presence of a complementary nucleic acid sequence in a given sample. Alternatively, the short nucleic acid sequences may be used as oligonucleotide primers to amplify or mutate a complementary nucleic acid sequence using PCR technology. These primers may also facilitate the amplification of related complementary nucleic acid sequences (*e.g.,* related sequences from other species).

Use of these probes or primers may greatly facilitate the identification of transgenic plants which contain the presently disclosed promoters and structural nucleic acid sequences. Such probes or primers may also be used to screen cDNA or genomic libraries for additional nucleic acid sequences related to or sharing homology with the presently disclosed promoters and structural nucleic acid sequences. The probes may also be PCR probes, which are nucleic acid molecules capable of initiating a polymerase activity while in a double-stranded structure with another nucleic acid.

A primer or probe is generally complementary to a portion of a nucleic acid sequence that is to be identified, amplified, or mutated and of sufficient length to form a stable and sequence-specific duplex molecule with its complement. The primer or probe preferably is about 10 to about 200 nucleotides long, more preferably is about 10 to about 100 nucleotides long, even more preferably is about 10 to about 50 nucleotides long, and most preferably is about 14 to about 30 nucleotides long.

The primer or probe may, for example without limitation, be prepared by direct chemical synthesis, by PCR (U.S. Patent Nos. 4,683,195 and 4,683,202), or by excising the nucleic acid specific fragment from a larger nucleic acid molecule. Various methods for determining the structure of PCR probes and PCR techniques exist in the art. Computer-generated searches using programs such as Primer3 (www-genome.wi.mit.

edu/cgi-bin/primer/primer3.cgi), STSPipeline (www-genome.wi.mit.edu/cgi-bin/www-STS_Pipeline), or GeneUp (Pesole *et al.*, *BioTechniques 25*:112-123, 1998), for example, can be used to identify potential PCR primers.

B. Protein and Peptide Molecules

Agents of the invention include proteins, peptide molecules, and fragments thereof encoded by nucleic acid agents of the invention. Preferred classes of protein and peptide molecules include: (1) GCPE proteins and peptide molecules; (2) GCPE proteins and peptide molecules derived from an organism having a MEP pathway; (3) GCPE proteins and peptide molecules derived from plants; and (4) GCPE proteins and peptide molecules derived from oilseed plants, including, but not limited to *Brassica campestris, Brassica napus,* canola, castor bean, coconut, cotton, crambe, linseed, maize, mustard, oil palm, peanut, rapeseed, rice, safflower, sesame, soybean, sunflower, and wheat.

Other preferred proteins are those proteins having an amino acid sequence: (1) selected from the group consisting of SEQ ID NOs: 4, 48, 49, and 50; (2) selected from the group consisting of SEQ ID NOs: 4, 48 and 49; (3) selected from the group consisting of SEQ ID NOs: 4 and 49; (4) of SEQ ID NO: 4; (5) of SEQ ID NO: 48; (6) of SEQ ID NO: 49; and (7) of SEQ ID NO: 50.

In another preferred aspect of the present invention the protein or peptide molecule is encoded by a nucleic acid agent of the invention, including, but not limited to a nucleic acid sequence that is selected from: (1) any of SEQ ID NOs: 1 through 3, 5 through 47, complements thereof, or fragments of these sequences; (2) the group consisting of SEQ ID NOs: 1, 2, complements thereof, and fragments of these sequences; (3) the group consisting of SEQ ID NOs: 1, 2, 3, complements thereof and fragments of these sequences; (4) the group consisting of SEQ ID NOs: 1, 2, 13 through 47, complements thereof and fragments of these sequences; (5) the group consisting of SEQ ID NOs: 5 through 12, complements thereof and fragments of these sequences; or (6) the

group consisting of SEQ ID NOs: 1 through 3, 5 through 47, complements thereof and fragments of these sequences.

Any of the nucleic acid agents of the invention may be linked with additional nucleic acid sequences to encode fusion proteins. The additional nucleic acid sequence preferably encodes at least one amino acid, peptide, or protein. Many possible fusion combinations exist. For instance, the fusion protein may provide a "tagged" epitope to facilitate detection of the fusion protein, such as GST, GFP, FLAG, or polyHIS. Such fusions preferably encode between 1 and about 50 amino acids, more preferably between about 5 and about 30 additional amino acids, and even more preferably between about 5 and about 20 amino acids.

Alternatively, the fusion may provide regulatory, enzymatic, cell signaling, or intercellular transport functions. For example, a sequence encoding a plastid transit peptide may be added to direct a fusion protein to the chloroplasts within seeds. Such fusion partners preferably encode between 1 and about 1000 additional amino acids, more preferably between about 5 and about 500 additional amino acids, and even more preferably between about 10 and about 250 amino acids.

The above-described protein or peptide molecules may be produced via chemical synthesis, or more preferably, by expression in a suitable bacterial or eukaryotic host. Suitable methods for expression are described by Sambrook *et al.*, *supra*, or similar texts. Fusion protein or peptide molecules of the invention are preferably produced via recombinant means. These proteins and peptide molecules may be derivatized to contain carbohydrate or other moieties (such as keyhole limpet hemocyanin, *etc.*).

Also contemplated are protein and peptide agents, including fragments and fusions thereof, in which conservative, non-essential or non-relevant amino acid residues have been added, replaced or deleted. A further particularly preferred class of protein is a GCPE protein, in which conservative, non-essential or non-relevant amino acid residues have been added, replaced or deleted. Computerized means for designing modifications

in protein structure are known in the art. *See, e.g.*, Dahiyat and Mayo, *Science* 278:82-87 (1997).

A protein of the invention can also be a homolog protein. In a preferred embodiment, the nucleic acid molecules of the present invention, complements thereof, and fragments of these sequences can be utilized to obtain such homologs. In another preferred embodiment, the homolog is selected from the group consisting of alfalfa, apple, *Arabidopsis,* banana, barley, *Brassica, Brassica campestris, Brassica napus*, broccoli, cabbage, canola, castor bean, chrysanthemum, citrus, coconut, coffee, cotton, crambe, cranberry, cucumber, Cuphea, dendrobium, dioscorea, eucalyptus, fescue, fir, garlic, gladiolus, grape, hordeum, lentils, lettuce, liliacea, linseed, maize, millet, muskmelon, mustard, oat, oil palm, oilseed rape, onion, an ornamental plant, papaya, pea, peanut, pepper, perennial ryegrass, *Phaseolus,* pine, poplar, potato, rapeseed (including Canola and High Erucic Acid varieties), rice, rye, safflower, sesame, sorghum, soybean, strawberry, sugarbeet, sugarcane, sunflower, tea, tomato, triticale, turf grasses, and wheat.

In a more preferred embodiment, the homolog is selected from *Brassica campestris, Brassica napus,* canola, castor bean, coconut, cotton, crambe, linseed, maize, mustard, oil palm, peanut, rapeseed (including Canola and High Erucic Acid varieties), rice, safflower, sesame, soybean, sunflower, and wheat, and in a particularly preferred embodiment from coconut, crambe, maize, oil palm, peanut, rapeseed (including Canola and High Erucic Acid varieties), safflower, sesame, soybean, and sunflower.

Agents of the invention include proteins comprising at least about a contiguous 10 amino acid region preferably comprising at least about a contiguous 20 amino acid region, even more preferably comprising at least about a contiguous 25, 35, 50, 75 or 100 amino acid region of a protein of the present invention. In another preferred embodiment, the proteins of the present invention include between about 10 and about 25 contiguous amino acid region, more preferably between about 20 and about 50 contiguous amino

acid region, and even more preferably between about 40 and about 80 contiguous amino acid region.

Due to the degeneracy of the genetic code, different nucleotide codons may be used to code for a particular amino acid. A host cell often displays a preferred pattern of codon usage. Nucleic acid sequences are preferably constructed to utilize the codon usage pattern of the particular host cell. This generally enhances the expression of the nucleic acid sequence in a transformed host cell. Any of the above described nucleic acid and amino acid sequences may be modified to reflect the preferred codon usage of a host cell or organism in which they are contained. Modification of a nucleic acid sequence for optimal codon usage in plants is described in U.S. Patent No. 5,689,052. Additional variations in the nucleic acid sequences may encode proteins having equivalent or superior characteristics when compared to the proteins from which they are engineered.

It is understood that certain amino acids may be substituted for other amino acids in a protein or peptide structure (and the nucleic acid sequence that codes for it) without appreciable change or loss of its biological utility or activity. For example, amino acid substitutions may be made without appreciable loss of interactive binding capacity in the antigen-binding regions of antibodies, or binding sites on substrate molecules. The modifications may result in either conservative or non-conservative changes in the amino acid sequence. The amino acid changes may be achieved by changing the codons of the nucleic acid sequence, according to the codons given in Table 1.

Table 1: Codon degeneracy of amino acids

Amino acid	One letter	Three letter	Codons
Alanine	A	Ala	GCA GCC GCG GCT
Cysteine	C	Cys	TGC TGT
Aspartic acid	D	Asp	GAC GAT
Glutamic acid	E	Glu	GAA GAG
Phenylalanine	F	Phe	TTC TTT
Glycine	G	Gly	GGA GGC GGG GGT

Amino acid	One letter	Three letter	Codons
Histidine	H	His	CAC CAT
Isoleucine	I	Ile	ATA ATC ATT
Lysine	K	Lys	AAA AAG
Leucine	L	Leu	TTA TTG CTA CTC CTG CTT
Methionine	M	Met	ATG
Asparagine	N	Asn	AAC AAT
Proline	P	Pro	CCA CCC CCG CCT
Glutamine	Q	Gln	CAA CAG
Arginine	R	Arg	AGA AGG CGA CGC CGG CGT
Serine	S	Ser	AGC AGT TCA TCC TCG TCT
Threonine	T	Thr	ACA ACC ACG ACT
Valine	V	Val	GTA GTC GTG GTT
Tryptophan	W	Trp	TGG
Tyrosine	Y	Tyr	TAC TAT

It is well known in the art that one or more amino acids in a native sequence can be substituted with other amino acid(s), the charge and polarity of which are similar to that of the native amino acid, *i.e.*, a conservative amino acid substitution, resulting in a silent change. Conservative substitutes for an amino acid within the native polypeptide sequence can be selected from other members of the class to which the amino acid belongs. Amino acids can be divided into the following four groups: (1) acidic (negatively charged) amino acids, such as aspartic acid and glutamic acid; (2) basic (positively charged) amino acids, such as arginine, histidine, and lysine; (3) neutral polar amino acids, such as glycine, serine, threonine, cysteine, cystine, tyrosine, asparagine, and glutamine; and (4) neutral nonpolar (hydrophobic) amino acids such as alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine.

In a further aspect of the present invention, nucleic acid molecules of the present invention can comprise sequences that differ from those encoding a protein or fragment thereof selected from the group consisting of SEQ ID NOs: 4 and 48 through 50 due to the fact that the different nucleic acid sequence encodes a protein having one or more conservative amino acid changes.

In a preferred aspect, biologically functional equivalents of the proteins or fragments thereof of the present invention can have about 10 or fewer conservative amino acid changes, more preferably about 7 or fewer conservative amino acid changes, and most preferably about 5 or fewer conservative amino acid changes. In a preferred embodiment, the protein has between about 5 and about 500 conservative changes, more preferably between about 10 and about 300 conservative changes, even more preferably between about 25 and about 150 conservative changes, and most preferably between about 5 and about 25 conservative changes or between 1 and about 5 conservative changes.

Non-conservative changes include additions, deletions, and substitutions that result in an altered amino acid sequence. In a preferred embodiment, the protein has between about 5 and about 500 non-conservative amino acid changes, more preferably between about 10 and about 300 non-conservative amino acid changes, even more preferably between about 25 and about 150 non-conservative amino acid changes, and most preferably between about 5 and about 25 non-conservative amino acid changes or between 1 and about 5 non-conservative changes.

In making such changes, the role of the hydropathic index of amino acids in conferring interactive biological function on a protein may be considered. *See* Kyte and Doolittle, *J. Mol. Biol. 157*:105-132 (1982). It is accepted that the relative hydropathic character of amino acids contributes to the secondary structure of the resultant protein, which in turn defines the interaction of the protein with other molecules, *e.g.*, enzymes, substrates, receptors, DNA, antibodies, antigens, *etc.* It is also understood in the art that the substitution of like amino acids may be made effectively on the basis of hydrophilicity, as the greatest local average hydrophilicity of a protein is known to correlate with a biological property of the protein. U.S. Patent No. 4,554,101.

Each amino acid has been assigned a hydropathic index and a hydrophilic value, as shown in Table 2.

Table 2: Amino Acid Hydropathic Indices and Hydrophilic Values

Amino acid	Hydropathic Index	Hydrophilic Value
Alanine	+1.8	-0.5
Cysteine	+2.5	-1.0
Aspartic acid	-3.5	+3.0 ±1
Glutamic acid	-3.5	+3.0 ±1
Phenylalanine	+2.8	-2.5
Glycine	-0.4	0
Histidine	-3.2	-0.5
Isoleucine	+4.5	-1.8
Lysine	-3.9	+3.0
Leucine	+3.8	-1.8
Methionine	+1.9	-1.3
Asparagine	-3.5	+0.2
Proline	-1.6	-0.5 ±1
Glutamine	-3.5	+0.2
Arginine	-4.5	+3.0
Serine	-0.8	+0.3
Threonine	-0.7	-0.4
Valine	+4.2	-1.5
Tryptophan	-0.9	-3.4
Tyrosine	-1.3	-2.3

It is known in the art that certain amino acids may be substituted by other amino acids having a similar hydropathic or hydrophilic index, score or value, and still result in a protein with similar biological activity, *i.e.*, still obtain a biologically functional protein. In making such changes, the substitution of amino acids whose hydropathic indices or hydrophilic values are within ±2 is preferred, those within ±1 are more preferred, and those within ±0.5 are most preferred.

As outlined above, amino acid substitutions are therefore based on the relative similarity of the amino acid side-chain substituents, for example, their hydrophobicity, hydrophilicity, charge, size, and the like. Exemplary substitutions which take various of the foregoing characteristics into consideration are well known to those of skill in the art and include: arginine and lysine; glutamate and aspartate; serine and threonine; glutamine and asparagine; and valine, leucine, and isoleucine.

These amino acid changes may be effected by mutating the nucleic acid sequence coding for the protein or peptide. Mutations to a nucleic acid sequence may be introduced in either a specific or random manner, both of which are well known to those of skill in the art of molecular biology. Mutations may include deletions, insertions, truncations, substitutions, fusions, shuffling of motif sequences, and the like. A myriad of site-directed mutagenesis techniques exist, typically using oligonucleotides to introduce mutations at specific locations in a structural nucleic acid sequence. Examples include single strand rescue, unique site elimination, nick protection, and PCR. Random or non-specific mutations may be generated by chemical agents (for a general review, *see* Singer and Kusmierek, *Ann. Rev. Biochem. 52*:655-693, 1982) such as nitrosoguanidine and 2-aminopurine; or by biological methods such as passage through mutator strains (Greener *et al., Mol. Biotechnol. 7*:189-195, 1997).

C. Recombinant Vectors and Constructs

Exogenous and/or heterologous genetic material may be transferred into a host cell by use of a vector or construct designed for such a purpose. Any of the nucleic acid sequences described above may be provided in a recombinant vector. The vector may be a linear or a closed circular plasmid. The vector system may be a single vector or plasmid or two or more vectors or plasmids that together contain the total DNA to be introduced into the genome of the host. Means for preparing recombinant vectors are well known in the art. Methods for making recombinant vectors particularly suited to plant transformation are described in U.S. Patent Nos.: 4,971,908, 4,940,835, 4,769,061 and 4,757,011.

Typical vectors useful for expression of nucleic acids in higher plants are well known in the art and include vectors derived from the tumor-inducing (Ti) plasmid of *Agrobacterium tumefaciens*. Other vector systems suitable for introducing transforming DNA into a host plant cell include, but are not limited to the pCaMVCN transfer control

vector, binary artificial chromosome (BIBAC) vectors (Hamilton *et al.*, Gene 200:107-116, 1997), and transfection with RNA viral vectors (Della-Cioppa *et al.*, *Ann. N.Y. Acad. Sci.* 792: 57-61, 1996). Additional vector systems also include plant selectable YAC vectors such as those described in Mullen *et al.*, *Molecular Breeding* 4:449-457 (1988).

A construct or vector may include a promoter, *e.g.*, a recombinant vector typically comprises, in a 5' to 3' orientation: a promoter to direct the transcription of a nucleic acid sequence of interest and a nucleic acid sequence of interest. Suitable promoters include, but are not limited to, those described herein. The recombinant vector may further comprise a 3' transcriptional terminator, a 3' polyadenylation signal, other untranslated nucleic acid sequences, transit and targeting nucleic acid sequences, selectable markers, enhancers, and operators, as desired.

The vector may be an autonomously replicating vector, *i.e.*, a vector that exists as an extrachromosomal entity, the replication of which is independent of chromosomal replication, *e.g.*, a plasmid, an extrachromosomal element, a minichromosome, or an artificial chromosome. The vector may contain any means for assuring self-replication. For autonomous replication, the vector may further comprise an origin of replication enabling the vector to replicate autonomously in the host cell in question. Alternatively, the vector may be one that, when introduced into the cell, is integrated into the genome and replicated together with the chromosome(s) into which it has been integrated. This integration may be the result of homologous or non-homologous recombination.

Integration of a vector or nucleic acid into the genome by homologous recombination, regardless of the host being considered, relies on the nucleic acid sequence of the vector. Typically, the vector contains nucleic acid sequences for directing integration by homologous recombination into the genome of the host. These nucleic acid sequences enable the vector to be integrated into the host cell genome at a precise location or locations in one or more chromosomes. To increase the likelihood of integration at a precise location, there should be preferably two nucleic acid sequences

that individually contain a sufficient number of nucleic acids, preferably about 400 bp to about 1500 bp, more preferably about 800 bp to about 1000 bp, which are highly homologous with the corresponding host cell target sequence. These nucleic acid sequences may be any sequence that is homologous with a host cell target sequence and, furthermore, may or may not encode proteins.

Vectors suitable for replication in mammalian cells may include viral replicons, or sequences that ensure integration of the appropriate sequences encoding HCV epitopes into the host genome. For example, another vector used to express foreign DNA is vaccinia virus. Such heterologous DNA is generally inserted into a gene that is non-essential to the virus, for example, the thymidine kinase gene (tk), which also provides a selectable marker. Expression of the HCV polypeptide then occurs in cells or animals that are infected with the live recombinant vaccinia virus.

In general, plasmid vectors containing replicon and control sequences that are derived from species compatible with the host cell are used in connection with bacterial hosts. The vector ordinarily carries a replication site, as well as marking sequences that are capable of providing phenotypic selection in transformed cells. For example, *E. coli* is typically transformed using pBR322, which contains genes for ampicillin and tetracycline resistance and thus provides easy means for identifying transformed cells. The pBR322 plasmid, or other microbial plasmid or phage, also generally contains, or is modified to contain, promoters that can be used by the microbial organism for expression of the selectable marker genes.

Promoters

Promoters used in the context of the present invention are selected on the basis of the cell type into which the vector will be inserted. Promoters that function in bacteria, yeast, and plants are all taught in the art. The promoters may also be selected on the basis of their regulatory features, e.g., enhancement of transcriptional activity,

inducibility, tissue specificity, and developmental stage-specificity. Additional promoters that may be utilized are described, for example, in U.S. Patent Nos. 5,378,619; 5,391,725; 5,428,147; 5,447,858; 5,608,144; 5,614,399; 5,633,441; 5,633,435; and 4,633,436.

Particularly preferred promoters in the recombinant vector include the nopaline synthase (*nos*) promoter; mannopine synthase (*mas*) promoter; octopine synthase (*ocs*) promoter; the cauliflower mosaic virus (CaMV) 19S and 35S promoters; the enhanced CaMV 35S promoter (eCaMV); the Figwort Mosaic Virus (FMV) 35S promoter; the light-inducible promoter from the small subunit of ribulose-1,5-bisphosphate carboxylase (ssRUBISCO); the EIF-4A promoter from tobacco; corn sucrose synthetase 1; corn alcohol dehydrogenase 1; corn light harvesting complex; corn heat shock protein; the chitinase promoter from *Arabidopsis*; the LTP (Lipid Transfer Protein) promoters from broccoli; petunia chalcone isomerase; bean glycine rich protein 1; potato patatin; the ubiquitin promoter from maize; the Adh promoter; the R gene complex promoter; and the actin promoter from rice.

The promoter is most preferably the *nos*, *ocs*, *mas*, CaMV19S, CaMV35S, eCaMV, ssRUBISCO, FMV, CaMV derived AS4, tobacco RB7, wheat POX1, tobacco EIF-4, lectin protein (Le1), or rice RC2 promoter. The promoter is preferably seed selective, tissue selective, constitutive, or inducible.

Often-used constitutive promoters include the CaMV 35S promoter, the eCaMV 35S promoter, the FMV promoter, the *mas* promoter, the *nos* promoter, and the *ocs* promoter, which is carried on tumor-inducing plasmids of *Agrobacterium tumefaciens*.

Useful inducible promoters include promoters induced by salicylic acid or polyacrylic acids (PR-1), induced by application of safeners (substituted benzenesulfonamide herbicides), heat-shock promoters, a nitrate-inducible promoter derived from the spinach nitrite reductase structural nucleic acid sequence, hormone-inducible promoters, and light-inducible promoters associated with the small subunit of RuBP carboxylase and LHCP families.

For the purposes of expression in specific tissues of the plant, such as the leaf, seed, root or stem, it is preferred that the promoters utilized have relatively high expression in these specific tissues or organs. Examples reported in the literature include the chloroplast glutamine synthetase GS2 promoter from pea, the chloroplast fructose-1,6-biphosphatase (FBPase) promoter from wheat, the nuclear photosynthetic ST-LS1 promoter from potato, the serine/threonine kinase (PAL) promoter and the glucoamylase (CHS) promoter from *A. thaliana*.

Also reported to be active in photosynthetically active tissues are the ribulose-1,5-bisphosphate carboxylase (RbcS) promoter from eastern larch (*Larix laricina*), the promoters for the *cab* genes of pine, wheat, spinach, and rice, the pyruvate orthophosphate dikinase (PPDK) promoter from maize, the promoter for the tobacco Lhcb1*2 gene, the *A. thaliana* SUC2 sucrose-H+ symporter promoter and the promoter for the thylakoid membrane proteins from spinach (*psaD, psaF, psaE, PC, FNR, atpC, atpD, cab, rbcS*). Other promoters for the chlorophyll a/b-binding proteins may also be utilized in the invention, such as the promoters for *LhcB* gene and *PsbP* gene from white mustard.

For the purpose of expression in sink tissues of the plant, such as the tuber of the potato plant, the fruit of tomato, or the seed of maize, wheat, rice and barley, it is preferred that the promoters utilized in the invention have relatively high expression in these specific tissues. A number of promoters for genes with tuber-specific or tuber-enhanced expression are known, including the class I patatin promoter, the promoter for the potato tuber ADPGPP genes, both the large and small subunits, the sucrose synthase promoter, the promoter for the major tuber proteins including the 22 kd protein complexes and protease inhibitors, the promoter for the granule-bound starch synthase gene (GBSS) and other class I and II patatins promoters.

Plant functional promoters useful for preferential expression in seeds include those from plant storage proteins and from proteins involved in fatty acid biosynthesis in

oilseeds. Examples of such promoters include the 5' regulatory regions from such genes as napin, phaseolin, zein, soybean trypsin inhibitor, ACP, stearoyl-ACP desaturase, soybean a' subunit of b-conglycinin (soy 7s), and oleosin. Further examples include the promoter for β-conglycinin and the lectin promoter from soybean. Seed-specific regulation is further discussed in EP 255 378.

Also included are promoters for the zeins, which are a group of storage proteins found in maize endosperm. Genomic clones for zein genes have been isolated and the promoters from these clones, including the 15 kD, 16 kD, 19 kD, 22 kD, 27 kD and genes, can also be used. Other promoters known to function, for example, in maize include the promoters for the following genes: *waxy, Brittle, Shrunken 2*, Branching enzymes I and II, starch synthases, debranching enzymes, oleosins, glutelins and sucrose synthases. A particularly preferred promoter for maize endosperm expression is the promoter for the glutelin gene from rice, more particularly the Osgt-1 promoter.

Examples of promoters suitable for expression in wheat include those promoters for the ADPglucose pyrosynthase (ADPGPP) subunits, the granule bound and other starch synthase, the branching and debranching enzymes, the embryogenesis-abundant proteins, the gliadins and the glutenins. Preferred promoters in rice include promoters for the ADPGPP subunits, the granule bound and other starch synthase, the branching enzymes, the debranching enzymes, sucrose synthases and the glutelins, and particularly preferred is the promoter for rice glutelin, Osgt-1. Preferred promoters for barley include those promoters for the ADPGPP subunits, the granule bound and other starch synthase, the branching enzymes, the debranching enzymes, sucrose synthases, the hordeins, the embryo globulins and the aleurone specific proteins.

Root specific promoters can also be used. An example of such a promoter is the promoter for the acid chitinase gene. Expression in root tissue can also be accomplished by utilizing the root specific subdomains of the CaMV35S promoter that have been

identified. Other root cell specific promoters include those reported by Conkling *et al.*. *Plant Physiol.* 93:1203-1211 (1990).

Examples of suitable promoters for use with filamentous fungi are obtained from the genes encoding *Aspergillus oryzae* TAKA amylase, *Rhizomucor miehei* aspartic proteinase, *Aspergillus niger* neutral alpha-amylase, *A. niger* acid stable alpha-amylase, *A. niger* or *A. awamori* glucoamylase (glaA), *Rhizomucor miehei* lipase, *Aspergillus oryzae* alkaline protease, *A. oryzae* triose phosphate isomerase, *Aspergillus nidulans* acetamidase and hybrids thereof. In a yeast host, preferred promoters include the *Saccharomyces cerevisiae* enolase (eno-1), the TAKA amylase, NA2-tpi (a hybrid of the promoters from the genes encoding *A. niger* neutral alpha-amylase and *A. oryzae* triose phosphate isomerase), glaA, *S. cerevisiae* GAL1 (galactokinase) and *S. cerevisiae* GPD (glyceraldehyde-3-phosphate dehydrogenase) promoters.

Suitable promoters for mammalian cells are also known in the art and include viral promoters, such as those from Simian Virus 40 (SV40), Rous sarcoma virus (RSV), adenovirus (ADV), cytomegalovirus (CMV), and bovine papilloma virus (BPV), as well as mammalian cell-derived promoters. Other preferred promoters include the hematopoietic stem cell-specific, *e.g.*, CD34, glucose-6-phosphotase, interleukin-1 alpha, CD11c integrin gene, GM-CSF, interleukin-5R alpha, interleukin-2, c-fos, h-ras, and DMD gene promoters.

Inducible promoters suitable for use with bacteria hosts include the -lactamase and lactose promoter systems, the arabinose promoter system, alkaline phosphatase, a tryptophan (trp) promoter system and hybrid promoters such as the tac promoter. However, other known bacterial inducible promoters are suitable. Promoters for use in bacterial systems also generally contain a Shine-Dalgarno sequence operably linked to the DNA encoding the polypeptide of interest.

Examples of suitable promoters for an algal host are light harvesting protein promoters obtained from photosynthetic organisms, *Chlorella* virus methyltransferase

promoters, CaMV 35 S promoter, PL promoter from bacteriophage λ, nopaline synthase promoter from the Ti plasmid of *A. tumefaciens*, and bacterial trp promoter.

Vectors for use with insect cells or insects may utilize a baculovirus transcriptional promoter including, *e.g.*, but not limited to the viral DNAs of *Autographa californica MNPV, Bombyx mori NPV, Trichoplusia ni MNPV, Rachiplusia ou MNPV or Galleria mellonella MNPV*, wherein the baculovirus transcriptional promoter is a baculovirus immediate-early gene IEl or IEN promoter; an immediate-early gene in combination with a baculovirus delayed-early gene promoter region selected from the group consisting of 39K and a *HindIII-k* fragment delayed-early gene; or a baculovirus late gene promoter.

Additional Nucleic Acid Sequences of Interest

The recombinant vector may also contain one or more additional nucleic acid sequences of interest. These additional nucleic acid sequences may generally be any sequences suitable for use in a recombinant vector. Such nucleic acid sequences include, without limitation, any of the nucleic acid sequences, and modified forms thereof, described above. The additional nucleic acid sequences may also be operably linked to any of the above described promoters. The one or more additional nucleic acid sequences may each be operably linked to separate promoters. Alternatively, the additional nucleic acid sequences may be operably linked to a single promoter (*i.e.* a single operon).

The additional nucleic acid sequences include, without limitation, those encoding seed storage proteins, fatty acid pathway enzymes, tocopherol biosynthetic enzymes, amino acid biosynthetic enzymes, and starch branching enzymes. Preferred seed storage proteins include zeins, 7S proteins, brazil nut protein, phenylalanine-free proteins, albumin, β-conglycinin, 11S proteins, alpha-hordothionin, arcelin seed storage proteins, lectins, and glutenin. Preferred fatty acid pathway enzymes include thioesterases and desaturases.

Preferred tocopherol biosynthetic enzymes include *tyrA, slr1736, ATPT2, dxs, dxr, GGPPS, HPPD, GMT, MT1, AANT1, slr1737*, and an antisense construct for homogentisic acid dioxygenase. Preferred additional nucleic acid sequences encode MEP pathway proteins including *ygbB, ygbP, ychB, yfgA, yfgB, dxs* and *dxr*. More preferred nucleic acid sequences include *yfgA* and *yfgB*, and still other preferred nucleic acid sequences include *ygbB*, *ychB* and *ygbP*. Preferred amino acid biosynthetic enzymes include anthranilate synthase, tryptophan decarboxylase, threonine decarboxylase, threonine deaminase, and aspartate kinase. Preferred starch branching enzymes include those set forth in U.S. Patent Nos. 6,232,122 and 6,147,279, and WO 97/22703.

Alternatively, the additional nucleic acid sequence may be designed to down-regulate a specific nucleic acid sequence. This is typically accomplished by operably linking the additional nucleic acid sequence, in an antisense orientation, with a promoter. One of ordinary skill in the art is familiar with such antisense technology. Any nucleic acid sequence may be negatively regulated in this manner. Preferable target nucleic acid sequences contain a low content of essential amino acids, yet are expressed at relatively high levels in particular tissues. For example, β-conglycinin and glycinin are expressed abundantly in seeds, but are nutritionally deficient with respect to essential amino acids. This antisense approach may also be used to effectively remove other undesirable proteins, such as antifeedants (*e.g.*, lectins), albumin, and allergens, from plant-derived foodstuffs.

Selectable and Screenable Markers

A vector or construct may also include a selectable marker. Selectable markers can also be used to select for plants or plant cells that contain the exogenous genetic material. Examples of such include, but are not limited to: a *neo* gene, which codes for kanamycin resistance and can be selected for using kanamycin, RptII, G418, hpt *etc.*; a bar gene, which codes for bialaphos resistance; a mutant EPSP synthase gene, *aad*A, which encodes glyphosate resistance; a nitrilase gene, which confers resistance to

bromoxynil; a mutant acetolactate synthase gene (ALS), which confers imidazolinone or sulphonylurea resistance, ALS, and a methotrexate resistant DHFR gene. The selectable marker is preferably GUS, green fluorescent protein (GFP), neomycin phosphotransferase II (*nptII*), luciferase (LUX), an antibiotic resistance coding sequence, or an herbicide (*e.g.*, glyphosate) resistance coding sequence. The selectable marker is most preferably a kanamycin, hygromycin, or herbicide resistance marker.

A vector or construct can also include a screenable marker. Screenable markers are useful to monitor expression. Exemplary screenable markers include: a β-glucuronidase or *uid*A gene (GUS), which encodes an enzyme for which various chromogenic substrates are known; an R-locus gene, which encodes a product that regulates the production of anthocyanin pigments (red color) in plant tissues; a β-lactamase gene, which encodes an enzyme for which various chromogenic substrates are known (*e.g.*, PADAC, a chromogenic cephalosporin); a luciferase gene; a *xyl*E gene, which encodes a catechol dioxygenase that can convert chromogenic catechols; an α-amylase gene; a tyrosinase gene, which encodes an enzyme capable of oxidizing tyrosine to DOPA and dopaquinone which in turn condenses to melanin; an α-galactosidase, which will turn a chromogenic α-galactose substrate.

Included within the terms "selectable or screenable marker genes" are also genes that encode a secretable marker whose secretion can be detected as a means of identifying or selecting for transformed cells. Examples include markers that encode a secretable antigen that can be identified by antibody interaction, or even secretable enzymes that can be detected catalytically. Secretable proteins fall into a number of classes, including small, diffusible proteins that are detectable, (*e.g.*, by ELISA), small active enzymes that are detectable in extracellular solution (*e.g.*, α-amylase, β-lactamase, phosphinothricin transferase), or proteins that are inserted or trapped in the cell wall (such as proteins which include a leader sequence such as that found in the expression unit of extension or

tobacco PR-S). Other possible selectable and/or screenable marker genes will be apparent to those of skill in the art.

Other Elements in the Recombinant Vector

Various cis-acting untranslated 5' and 3' regulatory sequences may be included in the recombinant nucleic acid vector to produce desirable regulatory features. A vector or construct may also include regulatory elements. Examples of such include the Adh intron 1, the sucrose synthase intron and the TMV omega element. These and other regulatory elements may be included when appropriate, and may be provided by the DNA sequence encoding the gene of interest or a convenient transcription termination region derived from a different gene source.

A 3′ non-translated region typically provides a transcriptional termination signal, and a polyadenylation signal that functions in plants to cause the addition of adenylate nucleotides to the 3′ end of the mRNA. Such 3' non-translated regions can be obtained from the 3' regions of the nopaline synthase (*nos*) coding sequence, a soybean 7Sα' storage protein coding sequence, the arcelin-5 coding sequence, the albumin coding sequence, and the pea ssRUBISCO E9 coding sequence. Particularly preferred 3' nucleic acid sequences include Arcelin-5 3', *nos* 3', E9 3', *adr*12 3', 7Sα' 3', 11S 3', USP 3', and albumin 3'.

Translational enhancers may also be incorporated as part of the recombinant vector, such as one or more 5' non-translated leader sequences that serve to enhance expression of the nucleic acid sequence. Such enhancer sequences may be desirable to increase or alter the translational efficiency of the resultant mRNA. Preferred 5' nucleic acid sequences include dSSU 5', PetHSP70 5', and GmHSP17.9 5'. Such sequences can be derived from the promoter selected to express the gene or can be specifically modified to increase translation of the mRNA. Such regions can also be obtained from viral RNAs, from suitable eukaryotic genes, or from a synthetic gene sequence. For a review

of optimizing expression of transgenes, see Koziel *et al.*, *Plant Mol. Biol.* 32:393-405 (1996).

The recombinant vector can further comprise a nucleic acid sequence encoding a transit peptide. This peptide may be useful for directing a protein to the extracellular space, a plastid, or to some other compartment inside or outside of the cell. (see, *e.g.*, EP 0218571; U.S. Patent Nos.: 4,940,835, 5,610,041, 5,618,988, and 6,107,060). The nucleic acid sequence in the recombinant vector may comprise introns. The introns may be heterologous with respect to the structural nucleic acid sequence. Preferred introns include the rice actin intron and the corn HSP70 intron.

A protein or fragment thereof encoding nucleic acid molecule of the invention may also be operably linked to a suitable leader sequence. A leader sequence is a nontranslated region of a mRNA that is important for translation by the host. The leader sequence is operably linked to the 5' terminus of the nucleic acid sequence encoding the protein or fragment thereof. A polyadenylation sequence may also be operably linked to the 3' terminus of the nucleic acid sequence of the invention. The polyadenylation sequence is a sequence that when transcribed is recognized by the host to add polyadenosine residues to transcribed mRNA.

A protein or fragment thereof encoding nucleic acid molecule of the invention may also be linked to a propeptide coding region. A propeptide is an amino acid sequence found at the amino terminus of a proprotein or proenzyme. Cleavage of the propeptide from the proprotein yields a mature biochemically active protein. The resulting polypeptide is known as a propolypeptide or proenzyme (or a zymogen in some cases). Propolypeptides are generally inactive and can be converted to mature active polypeptides by catalytic or autocatalytic cleavage of the propeptide from the propolypeptide or proenzyme.

The recombinant vectors can further comprise one or more sequences that encode one or more factors that are advantageous in the expression of the protein or peptide, for

example, an activator (*e.g.*, a trans-acting factor), a chaperone and a processing protease. An activator is a protein that activates transcription of a nucleic acid sequence encoding a polypeptide, a chaperone is a protein that assists another protein in folding properly, and a processing protease is a protease that cleaves a propeptide to generate a mature biochemically active polypeptide. The nucleic acids encoding one or more of these factors are preferably not operably linked to the nucleic acid encoding the protein or fragment thereof.

D. Transgenic Organisms, and Methods for Producing Same

One or more of the nucleic acid molecules or recombinant vectors of the invention may be used in plant transformation or transfection. For example, exogenous genetic material may be transferred into a plant cell and the plant cell regenerated into a whole, fertile or sterile plant. In a preferred embodiment, the exogenous genetic material includes a nucleic acid molecule of the present invention, preferably a nucleic acid molecule encoding a GCPE protein. In another preferred embodiment, the nucleic acid molecule has a sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, complements thereof and fragments of these sequences. Other preferred exogenous genetic material are nucleic acid molecules that encode a protein or fragment thereof having an amino acid sequence selected from the group consisting of SEQ ID NOs: 4 , and 48 through 50 or fragments thereof.

The invention is also directed to transgenic plants and transformed host cells that comprise, in a 5' to 3' orientation, a promoter operably linked to a heterologous nucleic acid sequence of interest. Additional nucleic acid sequences may be introduced into the plant or host cell, such as 3' transcriptional terminators, 3' polyadenylation signals, other untranslated nucleic acid sequences, transit or targeting sequences, selectable markers, enhancers, and operators. Preferred nucleic acid sequences of the present invention, including recombinant vectors, structural nucleic acid sequences, promoters, and other

regulatory elements, are described above in parts A through C of the Detailed Description. Another embodiment of the invention is directed to a method of producing such transgenic plants which generally comprises the steps of selecting a suitable plant, transforming the plant with a recombinant vector, and obtaining the transformed host cell.

A transformed host cell may generally be any cell which is compatible with the present invention. A transformed host plant or cell can be or derived from a plant, or from a cell or organism such as a mammalian cell, mammal, fish cell, fish, bird cell, bird, algae cell, algae, fungal cell, fungus, or bacterial cell. Preferred host and transformants include: fungal cells such as *Aspergillus*, yeasts, mammals, particularly bovine and porcine, insects, bacteria, and algae. Methods to transform such cells or organisms are known in the art. *See, e.g.,* EP 238023; Becker and Guarente, in: Abelson and Simon (eds.), *Guide to Yeast Genetics and Molecular Biology, Methods Enzymol.* 194: 182-187, Academic Press, Inc., New York; Bennett and LaSure (eds.), *More Gene Manipulations in Fungi*, Academic Press, CA, 1991; Hinnen *et al., PNAS 75*:1920, 1978; Ito *et al., J. Bacteriology 153*:163, 1983; Malardier *et al., Gene 78*:147-156, 1989; Yelton *et al., PNAS 81*:1470-1474, 1984.

Transfer of a nucleic acid that encodes a protein can result in expression or overexpression of that protein in a transformed cell, transgenic organism or transgenic plant. One or more of the proteins or fragments thereof encoded by nucleic acid molecules of the invention may be overexpressed in a transformed cell, transgenic organism or transgenic plant. Such expression or overexpression may be the result of transient or stable transfer of the exogenous genetic material.

In a preferred embodiment, expression or overexpression of a GCPE protein in a host provides in that host, relative to an untransformed host with a similar genetic background, an increased level of: (1) tocotrienols; (2) tocopherols; (3) -tocopherols; (4) -tocopherols; (5) isopentenyl diphosphate (IPP); (6) DMAPP; (7) a GCPE protein in a plastid; (8) isoprenoids; (9) carotenoids; (10) an isoprenoid-related compound selected

from the group consisting of IPP, DMAPP, and a GCPE protein; or (11) an isoprenoid compound selected from the group consisting of tocotrienols, tocopherols, terpenes, gibberellins, carotenoids, xanthophylls, -tocopherols, -tocopherols, IPP, DMAPP, and a GCPE protein.

The expressed protein may be detected using methods known in the art that are specific for the particular protein or fragment. These detection methods may include the use of specific antibodies, formation of an enzyme product, or disappearance of an enzyme substrate. For example, if the protein has enzymatic activity, an enzyme assay may be used. Alternatively, if polyclonal or monoclonal antibodies specific to the protein are available, immunoassays may be employed using the antibodies to the protein. The techniques of enzyme assay and immunoassay are well known to those skilled in the art.

The resulting protein may be recovered by methods known in the arts. For example, the protein may be recovered from the nutrient medium by procedures including, but not limited to, centrifugation, filtration, extraction, spray-drying, evaporation, or precipitation. The recovered protein may then be further purified by a variety of chromatographic procedures, *e.g.*, ion exchange chromatography, gel filtration chromatography, affinity chromatography, or the like. Reverse-phase high performance liquid chromatography (RP-HPLC), optionally employing hydrophobic RP-HPLC media, *e.g.*, silica gel, further purify the protein. Combinations of methods and means can also be employed to provide a substantially purified recombinant polypeptide or protein.

In another preferred embodiment, overexpression of the GCPE protein in a transgenic plant may provide tolerance to a variety of stresses, *e.g.,* oxidative stress tolerance such as to oxygen or ozone, UV tolerance, heat tolerance, drought tolerance, cold tolerance, or fungal/microbial pathogen tolerance.

As used herein in a preferred aspect, a tolerance or resistance to stress is determined by the ability of a plant, when challenged by a stress such as cold, to produce a plant having a higher yield than one without such tolerance or resistance to stress. In a

particularly preferred aspect of the present invention, the tolerance or resistance to stress is measured relative to a plant with a similar genetic background to the tolerant or resistance plant except that the plant expresses or overexpresses a GCPE protein.

Host Cells and Organisms

Preferred host plants and cells can be or be derived from alfalfa, apple, *Arabidopsis,* banana, barley, *Brassica, Brassica campestris, Brassica napus*, broccoli, cabbage, canola, castor bean, chrysanthemum, citrus, coconut, coffee, cotton, crambe, cranberry, cucumber, Cuphea, dendrobium, dioscorea, eucalyptus, fescue, fir, garlic, gladiolus, grape, hordeum, lentils, lettuce, liliacea, linseed, maize, millet, muskmelon, mustard, oat, oil palm, oilseed rape, onion, an ornamental plant, papaya, pea, peanut, pepper, perennial ryegrass, *Phaseolus,* pine, poplar, potato, rapeseed (including Canola and High Erucic Acid varieties), rice, rye, safflower, sesame, sorghum, soybean, strawberry, sugarbeet, sugarcane, sunflower, tea, tomato, triticale, turf grasses, and wheat.

In a more preferred embodiment, the host plants and cells are, or are derived from, *Brassica campestris, Brassica napus,* canola, castor bean, coconut, cotton, crambe, linseed, maize, mustard, oil palm, peanut, rapeseed (including Canola and High Erucic Acid varieties), rice, safflower, sesame, soybean, sunflower, and wheat, and in a particularly preferred embodiment from coconut, crambe, maize, oil palm, peanut, rapeseed (including Canola and High Erucic Acid varieties), safflower, sesame, soybean, and sunflower.

In another preferred embodiment, the plant or cell is or derived from canola. In another preferred embodiment, the plant or cell is or derived from *Brassica napus*. In a particularly preferred embodiment, the plant or cell is or derived from soybean. The soybean cell or plant is preferably a cell or plant of an elite soybean line.

Other preferred plants and plant host cells for use in the methods of the present invention include, but are not limited to Acacia, alfalfa, aneth, apple, apricot, artichoke,

arugula, asparagus, avocado, banana, barley, beet, blackberry, blueberry, broccoli, brussel sprouts, cabbage, canola, cantaloupe, carrot, cassava, cauliflower, celery, cherry, chicory, cilantro, citrus, clementines, coffee, corn, cotton, cucumber, Douglas fir, eggplant, endive, escarole, eucalyptus, fennel, figs, garlic, gourd, grape, grapefruit, honey dew, jicama, kiwifruit, lettuce, leeks, lemon, lime, Loblolly pine, mango, melon, nectarine, oat, oil palm, oilseed rape, okra, onion, orange, an ornamental plant, papaya, parsley, pea, peach, peanut, pear, pepper, persimmon, pine, pineapple, plantain, plum, pomegranate, poplar, potato, pumpkin, quince, radiata pine, radicchio, radish, raspberry, rice, rye, sorghum, Southern pine, soybean, spinach, squash, strawberry, sugarbeet, sugarcane, sunflower, sweet potato, sweetgum, tangerine, tea, tobacco, tomato, triticale, turf, turnip, a vine, watermelon, wheat, yams, and zucchini.

Mammalian cell lines available as hosts for expression are known in the art and include many immortalized cell lines available from the American Type Culture Collection (ATCC, Manassas, VA), such as HeLa cells, Chinese hamster ovary (CHO) cells, baby hamster kidney (BHK) cells and a number of other cell lines.

The fungal host cell may, for example, be a yeast cell, a fungi, or a filamentous fungal cell. In one embodiment, the fungal host cell is a yeast cell, and in a preferred embodiment, the yeast host cell is a cell of the species of *Candida, Kluyveromyces, Saccharomyces*, *Schizosaccharomyces, Pichia* and *Yarrowia.* In another embodiment, the fungal host cell is a filamentous fungal cell, and in a preferred embodiment, the filamentous fungal host cell is a cell of the species of *Acremonium, Aspergillus, Fusarium, Humicola, Myceliophthora, Mucor, Neurospora, Penicillium, Thielavia, Tolypocladium* and *Trichoderma.*

Suitable host bacteria include archaebacteria and eubacteria, especially eubacteria and most preferably *Enterobacteriaceae*. Examples of useful bacteria include *Escherichia, Enterobacter, Azotobacter, Erwinia, Bacillus, Pseudomonas, Klebsiella, Proteus, Salmonella, Serratia, Shigella, Rhizobia, Vitreoscilla and Paracoccus*. Suitable

E. coli hosts include *E. coli* W3110 (ATCC 27325), *E. coli* 294 (ATCC 31446), *E. coli B* and *E. coli* X1776 (ATCC 31537) (American Type Culture Collection, Manassas, Virginia). Mutant cells of any of the above-mentioned bacteria may also be employed. These hosts may be used with bacterial expression vectors such as *E. coli* cloning and expression vector Bluescript™ (Stratagene, La Jolla, CA); pIN vectors (Van Heeke and Schuster 1989), and pGEX vectors (Promega, Madison Wis.), which may be used to express foreign polypeptides as fusion proteins with glutathione S-transferase (GST).

Preferred insect host cells are derived from *Lepidopteran* insects such as *Spodoptera frugiperda* or *Trichoplusia ni*. The preferred *Spodoptera frugiperda* cell line is the cell line Sf9 (ATCC CRL 1711). Other insect cell systems, such as the silkworm *B. mori* can also be used. These host cells are preferably used in combination with Baculovirus expression vectors (BEVs), which are recombinant insect viruses in which the coding sequence for a chosen foreign gene has been inserted behind a baculovirus promoter in place of the viral gene, *e.g.*, polyhedrin (U.S. Patent No. 4,745,051).

Methods for Introducing Nucleic Acid Molecules into Organisms

Technology for introduction of nucleic acids into cells is well known to those of skill in the art. Common methods include chemical methods, microinjection, electroporation (U.S. Patent No. 5,384,253), particle acceleration, viral vectors, and receptor-mediated mechanisms. Fungal cells may be transformed by a process involving protoplast formation, transformation of the protoplasts and regeneration of the cell wall. The various techniques for transforming mammalian cells are also well known.

Algal cells may be transformed by a variety of known techniques, including but not limit to, microprojectile bombardment, protoplast fusion, electroporation, microinjection, and vigorous agitation in the presence of glass beads. Suitable procedures for transformation of green algal host cells are described in EP 108580. A suitable method of transforming cells of diatom *Phaeodactylum tricornutum* species is described

in WO 97/39106. Chlorophyll C-containing algae may be transformed using the procedures described in U.S. Patent No. 5,661,017.

Methods for introducing nucleic acids into plants are also well known. Suitable methods include bacterial infection (*e.g., Agrobacterium*), binary bacterial artificial chromosome vectors, direct delivery of nucleic acids (*e.g.,* via PEG-mediated transformation), desiccation/inhibition-mediated nucleic acid uptake, electroporation, agitation with silicon carbide fibers, and acceleration of nucleic acid coated particles, *etc.* (reviewed in Potrykus *et al.*, *Ann. Rev. Plant Physiol. Plant Mol. Biol. 42*:205, 1991). For example, electroporation has been used to transform maize protoplasts.

Alternatively, nucleic acids can be directly introduced into pollen by directly injecting a plant's reproductive organs. In another transformation technique, nucleic acids may also be injected into immature embryos. Plastids of higher plants can be stably transformed via particle gun delivery of DNA containing a selectable marker and targeting of the DNA to the plastid genome through homologous recombination (U.S. Patent Nos. 5,451,513 and 5,545,818).

Methods for transforming dicots, primarily by use of *Agrobacterium tumefaciens* and obtaining transgenic plants, have been published for cotton, soybean, *Brassica,* peanut, papaya, pea and *Arabidopsis thaliana. E.g.,* U.S. Patent Nos. 5,004,863, 5,159,135, 5,416,011 5,463,174, 5,518,908, and 5,569,834. The latter method for transforming *Arabidopsis thaliana* is commonly called "dipping" or vacuum infiltration or germplasm transformation. Transformation of monocotyledons using electroporation, particle bombardment and *Agrobacterium* has also been reported. Transformation and plant regeneration have been achieved in asparagus, barley, maize, oat, orchard grass, rice, rye, sugarcane, tall fescue, and wheat.

Transformation of plant protoplasts can be achieved using methods based on calcium phosphate precipitation, polyethylene glycol treatment, electroporation and combinations of these treatments. Application of these systems to different plant strains

depends upon the ability to regenerate that particular plant strain from protoplasts. Illustrative methods for the regeneration of cereals from protoplasts are described in Abdullah *et al., Biotechnology* 4:1087 (1986); Fujimura *et al., Plant Tissue Culture Letters* 2:74 (1985); Toriyama *et al., Theor Appl. Genet.* 205:34 (1986); and Yamada *et al., Plant Cell Rep.* 4:85 (1986).

To transform plant strains that cannot be successfully regenerated from protoplasts, other ways to introduce DNA into intact cells or tissues can be utilized. For example, cereals may be regenerated from immature embryos or explants. In addition, "particle gun" or high-velocity microprojectile technology can be utilized. Using the latter technology, DNA is carried through the cell wall and into the cytoplasm on the surface of small metal particles. The metal particles penetrate through several layers of cells and thus allow the transformation of cells within tissue explants. A particular advantage of microprojectile bombardment, in addition to it being an effective means of reproducibly transforming monocots, is that neither the isolation of protoplasts (Christou *et al., Plant Physiol. 87*:671-674, 1988), nor the susceptibility to *Agrobacterium* infection is required. *See also* Yang and Christou (eds.), *Particle Bombardment Technology for Gene Transfer*, Oxford Press, Oxford, England (1994).

An illustrative embodiment of a method for delivering DNA into maize cells by acceleration is a biolistics α-particle delivery system, which can be used to propel tungsten particles coated with DNA through a screen, such as a stainless steel or Nytex screen, onto a filter surface covered with corn cells cultured in suspension. Alternatively, immature embryos or other target cells may be arranged on solid culture medium. The screen disperses the tungsten nucleic acid particles so that they are not delivered to the recipient cells in large aggregates. A particle delivery system suitable for use with the invention is the helium acceleration PDS-1000/He gun, which is available from Bio-Rad Laboratories (Bio-Rad, Hercules, California).

Through the use of techniques set forth herein, one may obtain about 1000 or more loci of cells transiently expressing a marker gene. The number of cells in a focus which express the exogenous gene product 48 hours post-bombardment often ranges from one to ten, and average one to three.

In bombardment transformation, one may optimize the pre-bombardment culturing conditions and the bombardment parameters to yield the maximum numbers of stable transformants. Important physical parameters to adjust include physical parameters such as gap distance, flight distance, tissue distance and helium pressure. In addition, biological factors, such as the nature of transforming DNA (e.g., linearized DNA or intact supercoiled plasmids) and the manipulation of cells before and immediately after bombardment, may affect transformation optimization. It is believed that pre-bombardment manipulations are especially important for successful transformation of immature embryos. One may also minimize the trauma reduction factors by modifying conditions that influence the physiological state of the recipient cells and which may therefore influence transformation and integration efficiencies. For example, the osmotic state, tissue hydration and the subculture stage or cell cycle of the recipient cells may be adjusted for optimum transformation.

Agrobacterium-mediated transfer is a widely applicable system for introducing genes into plant cells because the DNA can be introduced into whole plant tissues, thereby bypassing the need for regeneration of an intact plant from a protoplast. Further, the integration of the Ti-DNA is a relatively precise process resulting in few rearrangements. The region of DNA to be transferred is defined by the border sequences and intervening DNA is usually inserted into the plant genome as described (Spielmann *et al.*, 1986).

Modern *Agrobacterium* transformation vectors are capable of replication in *E. coli* as well as *Agrobacterium*, allowing for convenient manipulations. Moreover, technological advances in vectors for *Agrobacterium*-mediated gene transfer have

improved the arrangement of genes and restriction sites in the vectors to facilitate construction of vectors capable of expressing various polypeptide coding genes. Available vectors have convenient multi-linker regions flanked by a promoter and a polyadenylation site for direct expression of inserted polypeptide coding genes and are suitable for present purposes. In addition, *Agrobacterium* containing both armed and disarmed Ti genes can be used for the transformations. In those plant strains where *Agrobacterium*-mediated transformation is efficient, it is the method of choice because of the facile and defined nature of the gene transfer.

A transgenic plant formed using *Agrobacterium* transformation methods typically contains a single gene on one chromosome. Such transgenic plants can be referred to as being heterozygous for the added gene. More preferred is a transgenic plant that is homozygous for the added structural gene; *i.e.*, a transgenic plant that contains two added genes, one gene at the same locus on each chromosome of a chromosome pair. A homozygous transgenic plant can be obtained by sexually mating (selfing) an independent segregant, transgenic plant that contains a single added gene, germinating some of the seed produced and analyzing the resulting plants produced for the gene of interest.

Transgenic Plants

Regeneration, development, and cultivation of plants from single plant protoplast transformants or various transformed explants is taught in the art, *e.g.,* by Weissbach and Weissbach (eds.), *Methods for Plant Molecular Biology,* Academic Press, Inc., San Diego, CA (1988); and Horsch *et al., Science 227*:1229-1231 (1985). There are a variety of methods for the regeneration of plants from plant tissue. The particular method of regeneration will depend on the starting plant tissue and the particular plant species to be regenerated.

Transformants are generally cultured in the presence of a selective media that selects for the successfully transformed cells and induces the regeneration of plant shoots. Such shoots are typically obtained within two to four months. Shoots are then transferred to an appropriate root-inducing medium containing the selective agent and an antibiotic to prevent bacterial growth. Many of the shoots will develop roots, which are then transplanted to soil or other media to allow the continued development of roots. The method, as outlined, will generally vary depending on the particular plant employed.

Preferably, the regenerated transgenic plants are self-pollinated to provide homozygous transgenic plants. Alternatively, pollen obtained from the regenerated transgenic plants may be crossed with seed-grown or non-transgenic plants, preferably plants of agronomically important lines. Conversely, pollen from seed-grown or non-transgenic plants may be used to pollinate the regenerated transgenic plants. A transgenic plant of the invention containing a desired polypeptide is cultivated using methods well-known to one skilled in the art.

A transgenic plant may pass along the nucleic acid sequence encoding the enhanced gene expression to its progeny. The transgenic plant is preferably homozygous for the nucleic acid encoding the enhanced gene expression and transmits that sequence to all of its offspring upon as a result of sexual reproduction. Progeny may be grown from seeds produced by the transgenic plant. These additional plants may then be self-pollinated to generate a true breeding line of plants.

It is also to be understood that two different transgenic plants can also be mated to produce offspring that contain two independently segregating, exogenous genes. Selfing of appropriate progeny can produce plants that are homozygous for both added, exogenous genes that encode a polypeptide of interest. Back-crossing to a parental plant and out-crossing with a non-transgenic plant are also contemplated, as is vegetative propagation.

The progeny from these plants are evaluated, among other things, for gene expression. The gene expression may be detected by several common methods such as western blotting, northern blotting, immunoprecipitation, and ELISA. Assays for gene expression based on the transient expression of cloned nucleic acid constructs have been developed by introducing the nucleic acid molecules into plant cells by polyethylene glycol treatment, electroporation, or particle bombardment. Transient expression systems may be used to functionally dissect gene constructs (*see generally*, Maliga *et al.*, *Methods in Plant Molecular Biology, A Laboratory Course Manual*, Cold Spring Harbor Press, Cold Spring Harbor, New York, 1995).

Any of the nucleic acid molecules of the invention may be introduced into a plant cell in a permanent or transient manner in combination with other genetic elements such as vectors, promoters, enhancers, *etc.* Further, any of the nucleic acid molecules of the invention may be introduced into a plant cell in a manner that allows for expression or overexpression of the protein or fragment thereof encoded by the nucleic acid molecule, for cosuppression of an endogenous protein, or for postranscriptional gene silencing of an endogenous transcript. In addition, the activity of a protein in a plant cell may be reduced or depressed by growing a transgenic plant cell containing a nucleic acid molecule whose non-transcribed strand encodes a protein or fragment thereof.

Cosuppression is the reduction in expression levels, usually at the level of RNA, of a particular endogenous gene or gene family by the expression of a homologous sense construct that is capable of transcribing mRNA of the same strandedness as the transcript of the endogenous gene. Cosuppression may result from stable transformation with a single copy nucleic acid molecule that is homologous to a nucleic acid sequence found with the cell or with multiple copies of a nucleic acid molecule that is homologous to a nucleic acid sequence found with the cell. Genes, even though different, linked to homologous promoters may result in the cosuppression of the linked genes.

Posttranscriptional gene silencing (PTGS) can result in virus immunity or gene silencing in plants. PTGS is induced by dsRNA and is mediated by an RNA-dependent RNA polymerase, present in the cytoplasm, that requires a dsRNA template. The dsRNA is formed by hybridization of complementary transgene mRNAs or complementary regions of the same transcript. Duplex formation can be accomplished by using transcripts from one sense gene and one antisense gene colocated in the plant genome, a single transcript that has self-complementarity, or sense and antisense transcripts from genes brought together by crossing. The dsRNA-dependent RNA polymerase makes a complementary strand from the transgene mRNA and RNAse molecules attach to this complementary strand (cRNA). These cRNA-RNase molecules hybridize to the endogene mRNA and cleave the single-stranded RNA adjacent to the hybrid. The cleaved single-stranded RNAs are further degraded by other host RNases because one will lack a capped 5' end and the other will lack a poly(A) tail. *See* Waterhouse *et al., PNAS 95*: 13959-13964 (1998).

Antisense approaches are a way of preventing or reducing gene function by targeting the genetic material. The objective of the antisense approach is to use a sequence complementary to the target gene to block its expression and create a mutant cell line or organism in which the level of a single chosen protein is selectively reduced or abolished. Antisense techniques have several advantages over other 'reverse genetic' approaches. The site of inactivation and its developmental effect can be manipulated by the choice of promoter for antisense genes or by the timing of external application or microinjection. Antisense can manipulate its specificity by selecting either unique regions of the target gene or regions where it shares homology to other related genes.

Under one embodiment, the process involves the introduction and expression of an antisense gene sequence. Such a sequence is one in which part or all of the normal gene sequences are placed under a promoter in inverted orientation so that the 'wrong' or complementary strand is transcribed into a noncoding antisense RNA that hybridizes with

the target mRNA and interferes with its expression. An antisense vector can be constructed by standard procedures and introduced into cells by transformation, transfection, electroporation, microinjection, infection, *etc.* The type of transformation and choice of vector will determine whether expression is transient or stable. The promoter used for the antisense gene may influence the level, timing, tissue, specificity, or inducibility of the antisense inhibition.

Feed, Meal, Protein and Oil Preparations

Plants or agents of the present invention can be utilized in methods, for example without limitation, to obtain a seed that expresses a *gcpE* nucleic acid molecule in that seed, to obtain a seed enhanced in a product of a *gcpE* gene, to obtain meal enhanced in a product of a *gcpE* gene, to obtain feedstock enhanced in a product of a *gcpE* gene, and to obtain oil enhanced in a product of a *gcpE* gene.

The present invention also provides for parts of the plants, particularly reproductive or storage parts, of the present invention. Plant parts, without limitation, include seed, endosperm, mesocarp, ovule and pollen. In a particularly preferred embodiment of the present invention, the plant part is a seed. In one embodiment the seed is a constituent of animal feed. In another embodiment, the plant part is a fruit, more preferably a fruit with enhanced shelf life. In another preferred embodiment, the fruit has increased levels of a tocopherol.

Plants utilized in such methods may be processed. A plant or plant part may be separated or isolated from other plant parts. A preferred plant part for this purpose is a seed. It is understood that even after separation or isolation from other plant parts, the isolated or separated plant part may be contaminated with other plant parts. In a preferred aspect, the separated plant part is greater than about 50% (w/w) of the separated material, more preferably, greater than about 75% (w/w) of the separated material, and even more preferably greater than about 90% (w/w) of the separated material. Plants or plant parts of

the present invention generated by such methods may be processed into products using known techniques.

Preferred products are meal, feedstock and oil. Methods to produce feed, meal, protein and oil preparations are known in the art. *See, e.g.*, U.S. Patents 4,957,748, 5,100,679, 5,219,596, 5,936,069, 6,005,076, 6,146,669, and 6,156,227. In a preferred embodiment, the protein preparation is a high protein preparation. Such a high protein preparation preferably has a protein content of greater than about 5% w/v, more preferably about 10% w/v, and even more preferably about 15% w/v.

In a preferred embodiment, the oil preparation is a high oil preparation with an oil content derived from a plant or part thereof of the present invention of greater than about 5% w/v, more preferably greater than about 10% w/v, and even more preferably greater than about 15% w/v. In a preferred embodiment the oil preparation is a liquid and of a volume greater than about 1, 5, 10 or 50 liters. The present invention provides for oil produced from plants of the present invention or generated by a method of the present invention. Such oil may be a minor or major component of any resultant product. Moreover, such oil may be blended with other oils.

In a preferred embodiment, the oil produced from plants of the present invention or generated by a method of the present invention constitutes greater than about 0.5%, 1%, 5%, 10%, 25%, 50%, 75% or 90% by volume or weight of the oil component of any product. In another embodiment, the oil preparation may be blended and can constitute greater than about 10%, 25%, 35%, 50% or 75% of the blend by volume. Oil produced from a plant of the present invention can be admixed with one or more organic solvents or petroleum distillates.

Seed containers

Seeds of the plants may be placed in a container. As used herein, a container is any object capable of holding such seeds. A container preferably contains greater than

about 500, 1,000, 5,000, or 25,000 seeds where at least about 10%, 25%, 50%, 75% or 100% of the seeds are derived from a plant of the present invention. The present invention also provides a container of over about 10,000, more preferably about 20,000, and even more preferably about 40,000 seeds where over about 10%, more preferably about 25%, more preferably 50% and even more preferably about 75% or 90% of the seeds are seeds derived from a plant of the present invention. The present invention also provides a container of over about 10 kg, more preferably about 25 kg, and even more preferably about 50 kg seeds where over about 10%, more preferably about 25%, more preferably about 50% and even more preferably about 75% or 90% of the seeds are seeds derived from a plant of the present invention.

E. Antibodies

One aspect of the invention concerns antibodies, single-chain antigen binding molecules, or other proteins that specifically bind to one or more of the protein or peptide molecules of the invention and their homologs, fusions or fragments. In a particularly preferred embodiment, the antibody specifically binds to a protein having the amino acid sequence set forth in SEQ ID NOs: 4, 48, 49 and 50, or an amino acid sequence encoded by a nucleic acid sequence selected from the group consisting of SEQ ID NOs: 1 through 3 and 5 through 47. Such antibodies may be used to quantitatively or qualitatively detect the protein or peptide molecules of the invention.

Nucleic acid molecules that encode all or part of the protein of the invention can be expressed, via recombinant means, to yield protein or peptides that can in turn be used to elicit antibodies that are capable of binding the expressed protein or peptide. Such antibodies may be used in immunoassays for that protein. Such protein-encoding molecules, or their fragments may be a "fusion" molecule (*i.e.*, a part of a larger nucleic acid molecule) such that, upon expression, a fusion protein is produced. It is understood

that any of the nucleic acid molecules of the invention may be expressed, via recombinant means, to yield proteins or peptides encoded by these nucleic acid molecules.

The antibodies that specifically bind proteins and protein fragments of the invention may be polyclonal or monoclonal and may comprise intact immunoglobulins, or antigen binding portions of immunoglobulins fragments (such as (F(ab'), $F(ab')_2$), or single-chain immunoglobulins producible, for example, via recombinant means. It is understood that practitioners are familiar with the standard resource materials that describe specific conditions and procedures for the construction, manipulation and isolation of antibodies (*see, e.g.*, Harlow and Lane, in: *Antibodies: A Laboratory Manual,* Cold Spring Harbor Press, Cold Spring Harbor, New York, 1988).

As discussed below, such antibody molecules or their fragments may be used for diagnostic purposes. Where the antibodies are intended for diagnostic purposes, it may be desirable to derivatize them, for example with a ligand group (such as biotin) or a detectable marker group (such as a fluorescent group, a radioisotope or an enzyme).

The ability to produce antibodies that bind the protein or peptide molecules of the invention permits the identification of mimetic compounds derived from those molecules. These mimetic compounds may contain a fragment of the protein or peptide or merely a structurally similar region and nonetheless exhibits an ability to specifically bind to antibodies directed against that compound.

Antibodies have been expressed in plants. Cytoplasmic expression of a scFv (single-chain Fv antibody) has been reported to delay infection by artichoke mottled crinkle virus. Transgenic plants that express antibodies directed against endogenous proteins may exhibit a physiological effect. For example, expressed anti-abscisic antibodies have been reported to result in a general perturbation of seed development. See, e.g., Hiatt *et al., Nature* 342:76-78 (1989); Conrad and Fielder, *Plant Mol. Biol.* 26:1023-1030 (1994); Philips *et al., EMBO J.* 16:4489-4496 (1997); Marion-Poll, *Trends in Plant Science* 2:447-448 (1997).

Antibodies that are catalytic may also be expressed in plants (abzymes). The principle behind abzymes is that because antibodies may be raised against many molecules, this recognition ability can be directed toward generating antibodies that bind transition states to force a chemical reaction forward. Persidas, *Nature Biotechnology* 15:1313-1315 (1997); Baca *et al., Ann. Rev. Biophys. Biomol. Struct.* 26:461-493 (1997). The catalytic abilities of abzymes may be enhanced by site directed mutagenesis. Examples of abzymes are, for example, set forth in U.S. Patent Nos. 5,658,753; 5,632,990; 5,631,137; 5,602,015; 5,559,538; 5,576,174; 5,500,358; 5,318,897; 5,298,409; 5,258,289; and 5,194,585. It is understood that any of the antibodies of the invention may be expressed in plants and that such expression can result in a physiological effect. It is also understood that any of the expressed antibodies may be catalytic.

F. Markers

Another subset of the nucleic acid molecules of the invention includes nucleic acid molecules that are markers. The markers can be used in a number of ways in the field of molecular genetics. Such markers include nucleic acid molecules SEQ ID NOs: 1 through 3 and 5 through 47 or complements thereof or fragments of either that can act as markers and other nucleic acid molecules of the present invention that can act as markers.

Genetic markers of the invention include "dominant" or "codominant" markers. "Codominant markers" reveal the presence of two or more alleles (two per diploid individual) at a locus. "Dominant markers" reveal the presence of only a single allele per locus. The presence of the dominant marker phenotype (*e.g.*, a band of DNA) is an indication that one allele is in either the homozygous or heterozygous condition. The absence of the dominant marker phenotype (*e.g.*, absence of a DNA band) is merely evidence that "some other" undefined allele is present. In the case of populations where individuals are predominantly homozygous and loci are predominately dimorphic, dominant and codominant markers can be equally valuable. As populations become more

heterozygous and multi-allelic, codominant markers often become more informative of the genotype than dominant markers. Marker molecules can be, for example, capable of detecting polymorphisms such as single nucleotide polymorphisms (SNPs).

The genomes of animals and plants naturally undergo spontaneous mutation in the course of their continuing evolution. A "polymorphism" is a variation or difference in the sequence of the gene or its flanking regions that arises in some of the members of a species. The variant sequence and the "original" sequence co-exist in the species' population. In some instances, such co-existence is in stable or quasi-stable equilibrium.

A polymorphism is thus said to be "allelic," in that, due to the existence of the polymorphism, some members of a species may have the original sequence (i.e., the original "allele") whereas other members may have the variant sequence (*i.e.*, the variant "allele"). In the simplest case, only one variant sequence may exist and the polymorphism is thus said to be di-allelic. In other cases, the species' population may contain multiple alleles and the polymorphism is termed tri-allelic, etc. A single gene may have multiple different unrelated polymorphisms. For example, it may have a di-allelic polymorphism at one site and a multi-allelic polymorphism at another site.

The variation that defines the polymorphism may range from a single nucleotide variation to the insertion or deletion of extended regions within a gene. In some cases, the DNA sequence variations are in regions of the genome that are characterized by short tandem repeats (STRs) that include tandem di- or tri-nucleotide repeated motifs of nucleotides. Polymorphisms characterized by such tandem repeats are referred to as "variable number tandem repeat" (VNTR) polymorphisms. VNTRs have been used in identity analysis (EP 370719; U.S. Patent Nos. 5,075,217 and 5,175,082; WO 91/14003).

The detection of polymorphic sites in a sample of DNA may be facilitated through the use of nucleic acid amplification methods. Such methods specifically increase the concentration of polynucleotides that span the polymorphic site, or include that site and

sequences located either distal or proximal to it. Such amplified molecules can be readily detected by gel electrophoresis or other means.

In an alternative embodiment, such polymorphisms can be detected through the use of a marker nucleic acid molecule that is physically linked to such polymorphism(s). For this purpose, marker nucleic acid molecules comprising a nucleotide sequence of a polynucleotide located within 1 mb of the polymorphism(s) and more preferably within 100kb of the polymorphism(s) and most preferably within 10kb of the polymorphism(s) can be employed. Alternatively, marker nucleic acid molecules comprising a nucleotide sequence of a polynucleotide located within 25 cM of the polymorphism(s) and more preferably within 15 cM of the polymorphism(s) and most preferably within 5 cM of the polymorphism(s) can be employed.

The identification of a polymorphism can be determined in a variety of ways. By correlating the presence or absence of it in a plant with the presence or absence of a phenotype, it is possible to predict the phenotype of that plant. If a polymorphism creates or destroys a restriction endonuclease cleavage site, or if it results in the loss or insertion of DNA (*e.g.*, a VNTR polymorphism), it will alter the size or profile of the DNA fragments that are generated by digestion with that restriction endonuclease. As such, organisms that possess a variant sequence can be distinguished from those having the original sequence by restriction fragment analysis. Polymorphisms that can be identified in this manner are termed "restriction fragment length polymorphisms" (RFLPs) (UK Patent Application 2135774; WO 90/13668; WO 90/11369).

Polymorphisms can also be identified by Single Strand Conformation Polymorphism (SSCP) analysis, random amplified polymorphic DNA (RAPD), and cleaveable amplified polymorphic sequences (CAPS). *See, e.g.*, Lee *et al., Anal. Biochem.* 205:289-293 (1992); Sarkar *et al., Genomics* 13:441-443 (1992); Williams *et al., Nucl. Acids Res.* 18:6531-6535 (1990); and Lyamichev *et al., Science* 260:778-783 (1993). It is understood that one or more of the nucleic acids of the invention, may be

utilized as markers or probes to detect polymorphisms by SSCP, RAPD or CAPS analysis.

Polymorphisms may also be found using a DNA fingerprinting technique called amplified fragment length polymorphism (AFLP), which is based on the selective PCR amplification of restriction fragments from a total digest of genomic DNA to profile that DNA. Vos *et al., Nucleic Acids Res.* 23:4407-4414 (1995). This method allows for the specific co-amplification of high numbers of restriction fragments, which can be visualized by PCR without knowledge of the nucleic acid sequence. It is understood that one or more of the nucleic acids of the invention may be utilized as markers or probes to detect polymorphisms by AFLP analysis or for fingerprinting RNA.

Single Nucleotide Polymorphisms (SNPs) generally occur at greater frequency than other polymorphic markers and are spaced with a greater uniformity throughout a genome than other reported forms of polymorphism. The greater frequency and uniformity of SNPs means that there is greater probability that such a polymorphism will be found near or in a genetic locus of interest than would be the case for other polymorphisms. SNPs are located in protein-coding regions and noncoding regions of a genome. Some of these SNPs may result in defective or variant protein expression (*e.g.*, as a result of mutations or defective splicing). Analysis (genotyping) of characterized SNPs can require only a plus/minus assay rather than a lengthy measurement, permitting easier automation.

SNPs can be characterized using any of a variety of methods. Such methods include the direct or indirect sequencing of the site, the use of restriction enzymes, enzymatic and chemical mismatch assays, allele-specific PCR, ligase chain reaction, single-strand conformation polymorphism analysis, single base primer extension (U.S. Patent Nos. 6,004,744 and 5,888,819), solid-phase ELISA-based oligonucleotide ligation assays, dideoxy fingerprinting, oligonucleotide fluorescence-quenching assays, 5'-nuclease allele-specific hybridization TaqMan™ assay, template-directed dye-terminator

incorporation (TDI) assay (Chen and Kwok, *Nucl. Acids Res.* 25:347-353, 1997), allele-specific molecular beacon assay (Tyagi *et al.*, *Nature Biotech.* 16: 49-53, 1998), PinPoint assay (Haff and Smirnov, *Genome Res.* 7: 378-388, 1997), dCAPS analysis (Neff *et al.*, *Plant J.* 14:387-392, 1998), pyrosequencing (Ronaghi *et al., Analytical Biochemistry* 267:65-71, 1999; WO 98/13523; WO 98/28440; and www.pyrosequencing.com), using mass spectrometry, *e.g.* the Masscode ™ system (WO 99/05319; WO 98/26095; WO 98/12355; WO 97/33000; WO 97/27331; www.rapigene.com; and U.S. Patent No. 5,965,363), invasive cleavage of oligonucleotide probes, and using high density oligonucleotide arrays (Hacia *et al.*, Nature Genetics 22:164-167; www.affymetrix.com).

Polymorphisms may also be detected using allele-specific oligonucleotides (ASO), which, can be for example, used in combination with hybridization based technology including Southern, northern, and dot blot hybridizations, reverse dot blot hybridizations and hybridizations performed on microarray and related technology.

The stringency of hybridization for polymorphism detection is highly dependent upon a variety of factors, including length of the allele-specific oligonucleotide, sequence composition, degree of complementarity (*i.e.* presence or absence of base mismatches), concentration of salts and other factors such as formamide, and temperature. These factors are important both during the hybridization itself and during subsequent washes performed to remove target polynucleotide that is not specifically hybridized. In practice, the conditions of the final, most stringent wash are most critical. In addition, the amount of target polynucleotide that is able to hybridize to the allele-specific oligonucleotide is also governed by such factors as the concentration of both the ASO and the target polynucleotide, the presence and concentration of factors that act to "tie up" water molecules, so as to effectively concentrate the reagents (*e.g.*, PEG, dextran, dextran sulfate, *etc.*), whether the nucleic acids are immobilized or in solution, and the duration of hybridization and washing steps.

Hybridizations are preferably performed below the melting temperature (T_m) of the ASO. The closer the hybridization and/or washing step is to the T_m, the higher the stringency. T_m for an oligonucleotide may be approximated, for example, according to the following formula: $T_m = 81.5 + 16.6 \times (\log10[Na+]) + 0.41 \times (\%G+C) - 675/n$; where [Na+] is the molar salt concentration of Na+ or any other suitable cation and n = number of bases in the oligonucleotide. Other formulas for approximating T_m are available and are known to those of ordinary skill in the art.

Stringency is preferably adjusted so as to allow a given ASO to differentially hybridize to a target polynucleotide of the correct allele and a target polynucleotide of the incorrect allele. Preferably, there will be at least a two-fold differential between the signal produced by the ASO hybridizing to a target polynucleotide of the correct allele and the level of the signal produced by the ASO cross-hybridizing to a target polynucleotide of the incorrect allele (*e.g.*, an ASO specific for a mutant allele cross-hybridizing to a wild-type allele). In more preferred embodiments of the present invention, there is at least a five-fold signal differential. In highly preferred embodiments of the present invention, there is at least an order of magnitude signal differential between the ASO hybridizing to a target polynucleotide of the correct allele and the level of the signal produced by the ASO cross-hybridizing to a target polynucleotide of the incorrect allele. While certain methods for detecting polymorphisms are described herein, other detection methodologies may be utilized.

The present invention includes and provides a method for detecting a polymorphism in a plant whose presence is predictive of a mutation affecting a level or pattern of a protein comprising: (A) incubating under conditions permitting nucleic acid hybridization: (i) a marker nucleic acid molecule having a nucleic acid sequence that hybridizes to a sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof; and (ii) a complementary nucleic acid molecule obtained from a sample, wherein nucleic acid hybridization between the marker nucleic

acid molecule and the complementary nucleic acid molecule permits the detection of a polymorphism; (B) permitting hybridization between the marker nucleic acid molecule and the complementary nucleic acid molecule; and (C) detecting the presence of the polymorphism, wherein the detection of the polymorphism is predictive of the mutation.

The present invention includes and provides a method of determining a degree of association between a polymorphism and a plant trait comprising: (A) hybridizing a nucleic acid molecule specific for the polymorphism to genetic material of a plant, wherein the nucleic acid molecule has a sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, complements thereof, and fragments of these sequences; and (B) calculating the degree of association between the polymorphism and the plant trait.

The present invention includes and provides a method of isolating a nucleic acid that encodes a protein or fragment thereof comprising: (A) incubating under conditions permitting nucleic acid hybridization: (i) a first nucleic acid molecule comprising a sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, complements thereof, and fragments of these sequences; and (ii) a complementary second nucleic acid molecule obtained from a plant cell or plant tissue; (B) permitting hybridization between the first nucleic acid molecule and the second nucleic acid molecule obtained from the plant cell or plant tissue; and (C) isolating the second nucleic acid molecule.

G. Plant Breeding

Plants of the present invention can be part of or generated from a breeding program. The choice of breeding method depends on the mode of plant reproduction, the heritability of the trait(s) being improved, and the type of cultivar used commercially (*e.g.*, F_1 hybrid cultivar, pureline cultivar, etc). Selected, non-limiting approaches, for breeding the plants of the present invention are set forth below. A breeding program can

be enhanced using marker assisted selection of the progeny of any cross. It is further understood that any commercial and non-commercial cultivars can be utilized in a breeding program. Factors such as, for example, emergence vigor, vegetative vigor, stress tolerance, disease resistance, branching, flowering, seed set, seed size, seed density, standability, and threshability etc. will generally dictate the choice.

For highly heritable traits, a choice of superior individual plants evaluated at a single location will be effective, whereas for traits with low heritability, selection should be based on mean values obtained from replicated evaluations of families of related plants. Popular selection methods commonly include pedigree selection, modified pedigree selection, mass selection, and recurrent selection. In a preferred embodiment a backcross or recurrent breeding program is undertaken.

The complexity of inheritance influences choice of the breeding method. Backcross breeding can be used to transfer one or a few favorable genes for a highly heritable trait into a desirable cultivar. This approach has been used extensively for breeding disease-resistant cultivars. Various recurrent selection techniques are used to improve quantitatively inherited traits controlled by numerous genes. The use of recurrent selection in self-pollinating crops depends on the ease of pollination, the frequency of successful hybrids from each pollination, and the number of hybrid offspring from each successful cross.

Breeding lines can be tested and compared to appropriate standards in environments representative of the commercial target area(s) for two or more generations. The best lines are candidates for new commercial cultivars; those still deficient in traits may be used as parents to produce new populations for further selection.

One method of identifying a superior plant is to observe its performance relative to other experimental plants and to a widely grown standard cultivar. If a single observation is inconclusive, replicated observations can provide a better estimate of its

genetic worth. A breeder can select and cross two or more parental lines, followed by repeated selfing and selection, producing many new genetic combinations.

The development of new cultivars requires the development and selection of varieties, the crossing of these varieties and the selection of superior hybrid crosses. The hybrid seed can be produced by manual crosses between selected male-fertile parents or by using male sterility systems. Hybrids are selected for certain single gene traits such as pod color, flower color, seed yield, pubescence color, or herbicide resistance, which indicate that the seed is truly a hybrid. Additional data on parental lines, as well as the phenotype of the hybrid, influence the breeder's decision whether to continue with the specific hybrid cross.

Pedigree breeding and recurrent selection breeding methods can be used to develop cultivars from breeding populations. Breeding programs combine desirable traits from two or more cultivars or various broad-based sources into breeding pools from which cultivars are developed by selfing and selection of desired phenotypes. New cultivars can be evaluated to determine which have commercial potential.

Pedigree breeding is used commonly for the improvement of self-pollinating crops. Two parents who possess favorable, complementary traits are crossed to produce an F_1. An F_2 population is produced by selfing one or several F_1's. Selection of the best individuals from the best families is carried out. Replicated testing of families can begin in the F_4 generation to improve the effectiveness of selection for traits with low heritability. At an advanced stage of inbreeding (*i.e.*, F_6 and F_7), the best lines or mixtures of phenotypically similar lines are tested for potential release as new cultivars.

Backcross breeding has been used to transfer genes for a simply inherited, highly heritable trait into a desirable homozygous cultivar or inbred line, which is the recurrent parent. The source of the trait to be transferred is called the donor parent. The resulting plant is expected to have the attributes of the recurrent parent (*e.g.*, cultivar) and the desirable trait transferred from the donor parent. After the initial cross, individuals

possessing the phenotype of the donor parent are selected and repeatedly crossed (backcrossed) to the recurrent parent. The resulting parent is expected to have the attributes of the recurrent parent (*e.g.*, cultivar) and the desirable trait transferred from the donor parent.

The single-seed descent procedure in the strict sense refers to planting a segregating population, harvesting a sample of one seed per plant, and using the one-seed sample to plant the next generation. When the population has been advanced from the F_2 to the desired level of inbreeding, the plants from which lines are derived will each trace to different F_2 individuals. The number of plants in a population declines each generation due to failure of some seeds to germinate or some plants to produce at least one seed. As a result, not all of the F_2 plants originally sampled in the population will be represented by a progeny when generation advance is completed.

In a multiple-seed procedure, breeders commonly harvest one or more pods from each plant in a population and thresh them together to form a bulk. Part of the bulk is used to plant the next generation and part is put in reserve. The procedure has been referred to as modified single-seed descent or the pod-bulk technique. The multiple-seed procedure has been used to save labor at harvest. It is considerably faster to thresh pods with a machine than to remove one seed from each by hand for the single-seed procedure. The multiple-seed procedure also makes it possible to plant the same number of seed of a population each generation of inbreeding.

Descriptions of other breeding methods that are commonly used for different traits and crops can be found in one of several reference books (*e.g.*, Fehr, *Principles of Cultivar Development*, Vol. 1 (1987).

A transgenic plant of the present invention may also be reproduced using apomixis. Apomixis is a genetically controlled method of reproduction in plants where the embryo is formed without union of an egg and a sperm. There are three basic types of apomictic reproduction: 1) apospory where the embryo develops from a chromosomally

unreduced egg in an embryo sac derived from the nucleus, 2) diplospory where the embryo develops from an unreduced egg in an embryo sac derived from the megaspore mother cell, and 3) adventitious embryony where the embryo develops directly from a somatic cell. In most forms of apomixis, pseudogamy or fertilization of the polar nuclei to produce endosperm is necessary for seed viability. In apospory, a nurse cultivar can be used as a pollen source for endosperm formation in seeds. The nurse cultivar does not affect the genetics of the aposporous apomictic cultivar because the unreduced egg of the cultivar develops parthenogenetically, but makes possible endosperm production. Apomixis is economically important, especially in transgenic plants, because it causes any genotype, no matter how heterozygous, to breed true. Thus, with apomictic reproduction, heterozygous transgenic plants can maintain their genetic fidelity throughout repeated life cycles. Methods for the production of apomictic plants are known in the art. *See, e.g.*, U.S. Patent No. 5,811,636.

Requirements for marker-assisted selection in a plant breeding program are: (1) the marker(s) should co-segregate or be closely linked with the desired trait; (2) an efficient means of screening large populations for the molecular marker(s) should be available; and (3) the screening technique should have high reproducibility across laboratories and preferably be economical to use and be user-friendly.

The genetic linkage of marker molecules can be established by a gene mapping model such as, without limitation, the flanking marker model reported by Lander and Botstein, *Genetics* 121:185-199 (1989), and the interval mapping model, based on maximum likelihood methods described by Lander and Botstein, and implemented in the software package MAPMAKER/QTL (Lincoln and Lander, *Mapping Genes Controlling Quantitative Traits Using MAPMAKER/QTL,* Whitehead Institute for Biomedical Research, Massachusetts, 1990). Additional software includes Qgene, Version 2.23 (1996), Department of Plant Breeding and Biometry, 266 Emerson Hall, Cornell University, Ithaca, NY). Use of Qgene software is a particularly preferred approach.

A maximum likelihood estimate (MLE) for the presence of a marker is calculated, together with an MLE assuming no QTL effect, to avoid false positives. A $\log_{10}$ of an odds ratio (LOD) is then calculated as: LOD = $\log_{10}$(MLE for the presence of a QTL/MLE given no linked QTL).

The LOD score essentially indicates how much more likely the data are to have arisen assuming the presence of a QTL than in its absence. The LOD threshold value for avoiding a false positive with a given confidence, say 95%, depends on the number of markers and the length of the genome. Graphs indicating LOD thresholds are set forth in Lander and Botstein, *supra*, and further described by Arús and Moreno-González, Plant Breeding, (Hayward *et al.*, eds.) Chapman & Hall, London, pp. 314-331 (1993).

In a preferred embodiment of the present invention the nucleic acid marker exhibits a LOD score of greater than about 2.0, more preferably about 2.5, even more preferably greater than about 3.0 or 4.0 with the trait or phenotype of interest. In a preferred embodiment, the trait of interest is altered tocopherol levels or compositions.

Additional models can be used. Many modifications and alternative approaches to interval mapping have been reported, including the use non-parametric methods. Kruglyak and Lander, *Genetics* 139:1421-1428 (1995). Multiple regression methods or models can be also be used, in which the trait is regressed on a large number of markers. Weber and Wricke, *Advances in Plant Breeding*, Blackwell, Berlin (1994). Procedures may combine interval mapping with regression analysis, whereby the phenotype is regressed onto a single putative QTL at a given marker interval and at the same time onto a number of markers that serve as 'cofactors.' Generally, the use of cofactors reduces the bias and sampling error of the estimated QTL positions, thereby improving the precision and efficiency of QTL mapping. Zeng, *Genetics* 136:1457-1468 (1994). These models can be extended to multi-environment experiments to analyze genotype-environment interactions. Jansen *et al., Theo. Appl. Genet.* 91:33-37 (1995).

It is understood that one or more of the nucleic acid molecules of the invention may be used as molecular markers. It is also understood that one or more of the protein molecules of the invention may be used as molecular markers.

In a preferred embodiment, the polymorphism is present and screened for in a mapping population, *e.g.* a collection of plants capable of being used with markers such as polymorphic markers to map genetic position of traits. The choice of appropriate mapping population often depends on the type of marker systems employed. Consideration must be given to the source of parents (adapted vs. exotic) used in the mapping population. Chromosome pairing and recombination rates can be severely disturbed (suppressed) in wide crosses (adapted x exotic) and generally yield greatly reduced linkage distances. Wide crosses will usually provide segregating populations with a relatively large number of polymorphisms when compared to progeny in a narrow cross (adapted x adapted).

An F_2 population is the first generation of selfing (self-pollinating) after the hybrid seed is produced. Usually a single F_1 plant is selfed to generate a population segregating for all the genes in Mendelian (1:2:1) pattern. Maximum genetic information is obtained from a completely classified F_2 population using a codominant marker system (Mather, 1938). In the case of dominant markers, progeny tests (*e.g.*, F_3, BCF_2) are required to identify the heterozygotes, in order to classify the population. However, this procedure is often prohibitive because of the cost and time involved in progeny testing. Progeny testing of F_2 individuals is often used in map construction where phenotypes do not consistently reflect genotype (*e.g.* disease resistance) or where trait expression is controlled by a QTL. Segregation data from progeny test populations *e.g.* F_3 or BCF_2) can be used in map construction. Marker-assisted selection can then be applied to cross progeny based on marker-trait map associations (F_2, F_3), where linkage groups have not been completely disassociated by recombination events (*i.e.*, maximum disequilibrium).

Recombinant inbred lines (RIL) (genetically related lines; usually $>F_5$, developed from continuously selfing F_2 lines towards homozygosity) can be used as a mapping population. Information obtained from dominant markers can be maximized by using RIL because all loci are homozygous or nearly so. Under conditions of tight linkage (*i.e.*, about <10% recombination), dominant and co-dominant markers evaluated in RIL populations provide more information per individual than either marker type in backcross populations. However, as the distance between markers becomes larger (*i.e.*, loci become more independent), the information in RIL populations decreases dramatically when compared to codominant markers.

Backcross populations (*e.g.*, generated from a cross between a successful variety (recurrent parent) and another variety (donor parent) carrying a trait not present in the former) can be utilized as a mapping population. A series of backcrosses to the recurrent parent can be made to recover most of its desirable traits. Thus a population is created consisting of individuals nearly like the recurrent parent but each individual carries varying amounts or mosaic of genomic regions from the donor parent. Backcross populations can be useful for mapping dominant markers if all loci in the recurrent parent are homozygous and the donor and recurrent parent have contrasting polymorphic marker alleles.

Information obtained from backcross populations using either codominant or dominant markers is less than that obtained from F_2 populations because one, rather than two, recombinant gamete is sampled per plant. Backcross populations, however, are more informative (at low marker saturation) when compared to RILs as the distance between linked loci increases in RIL populations (*i.e.* about .15% recombination). Increased recombination can be beneficial for resolution of tight linkages, but may be undesirable in the construction of maps with low marker saturation.

Near-isogenic lines (NIL) (created by many backcrosses to produce a collection of individuals that is nearly identical in genetic composition except for the trait or genomic

region under interrogation) can be used as a mapping population. In mapping with NILs, only a portion of the polymorphic loci is expected to map to a selected region.

Bulk segregant analysis (BSA) is a method developed for the rapid identification of linkage between markers and traits of interest (Michelmore *et al.*, *PNAS* 88:9828-9832 (1991). In BSA, two bulked DNA samples are drawn from a segregating population originating from a single cross. These bulks contain individuals that are identical for a particular trait (resistant or susceptible to particular disease) or genomic region but arbitrary at unlinked regions (*i.e.* heterozygous). Regions unlinked to the target region will not differ between the bulked samples of many individuals in BSA.

H. Determining the Level of Expression Response

In an aspect of the present invention, one or more of the nucleic molecules of the present invention are used to determine the level (*i.e.*, the concentration of mRNA in a sample, *etc.*) or pattern (*i.e.*, the kinetics of expression, rate of decomposition, stability profile, *etc.*) of the expression of a protein encoded in part or whole by one or more of the nucleic acid molecule of the present invention (collectively, the "Expression Response" of a cell or tissue).

As used herein, the Expression Response manifested by a cell or tissue is said to be "altered" if it differs from the Expression Response of cells or tissues of plants not exhibiting the phenotype. To determine whether a Expression Response is altered, the Expression Response manifested by the cell or tissue of the plant exhibiting the phenotype is compared with that of a similar cell or tissue sample of a plant not exhibiting the phenotype. As will be appreciated, it is not necessary to re-determine the Expression Response of the cell or tissue sample of plants not exhibiting the phenotype each time such a comparison is made; rather, the Expression Response of a particular plant may be compared with previously obtained values of normal plants.

A change in genotype or phenotype may be transient or permanent. Also as used herein, a tissue sample is any sample that comprises more than one cell. In a preferred aspect, a tissue sample comprises cells that share a common characteristic (e.g. derived from root, seed, flower, leaf, stem or pollen etc.).

In one aspect of the present invention, an evaluation can be conducted to determine whether a particular mRNA molecule is present. One or more of the nucleic acid molecules of the present invention are utilized to detect the presence or quantity of the mRNA species. Such molecules are then incubated with cell or tissue extracts of a plant under conditions sufficient to permit nucleic acid hybridization. The detection of double-stranded probe-mRNA hybrid molecules is indicative of the presence of the mRNA; the amount of such hybrid formed is proportional to the amount of mRNA. Thus, such probes may be used to ascertain the level and extent of the mRNA production in a plant's cells or tissues. Such nucleic acid hybridization may be conducted under quantitative conditions (thereby providing a numerical value of the amount of the mRNA present). Alternatively, the assay may be conducted as a qualitative assay that indicates either that the mRNA is present, or that its level exceeds a user set, predefined value.

A number of methods can be used to compare the expression response between two or more samples of cells or tissue. These methods include hybridization assays, such as northerns, RNAse protection assays, and *in situ* hybridization. Alternatively, the methods include PCR-type assays. In a preferred method, the expression response is compared by hybridizing nucleic acids from the two or more samples to an array of nucleic acids. The array contains a plurality of suspected sequences known or suspected of being present in the cells or tissue of the samples.

An advantage of *in situ* hybridization over more other techniques for the detection of nucleic acids is that it allows an investigator to determine the precise spatial population. *In situ* hybridization may be used to measure the steady-state level of RNA

accumulation. A number of protocols have been devised for *in situ* hybridization, each with tissue preparation, hybridization and washing conditions.

In situ hybridization also allows for the localization of proteins within a tissue or cell. It is understood that one or more of the molecules of the invention, preferably one or more of the nucleic acid molecules or fragments thereof of the invention or one or more of the antibodies of the invention may be utilized to detect the level or pattern of a protein or mRNA thereof by *in situ* hybridization.

Fluorescent *in situ* hybridization allows the localization of a particular DNA sequence along a chromosome, which is useful, among other uses, for gene mapping, following chromosomes in hybrid lines, or detecting chromosomes with translocations, transversions or deletions. *In situ* hybridization has been used to identify chromosomes in several plant species. It is understood that the nucleic acid molecules of the invention may be used as probes or markers to localize sequences along a chromosome.

Another method to localize the expression of a molecule is tissue printing. Tissue printing provides a way to screen, at the same time on the same membrane many tissue sections from different plants or different developmental stages. *See, e.g.*, Barres *et al., Neuron* 5:527-544 (1990); Cassab and Varner, *J. Cell. Biol.* 105:2581-2588 (1987); Harris and Chrispeels, *Plant Physiol.* 56:292-299 (1975); Reid and Pont-Lezica, *Tissue Printing: Tools for the Study of Anatomy, Histochemistry and Gene Expression*, Academic Press, New York, New York (1992); Reid *et al., Plant Physiol.* 93:160-165 (1990); Spruce *et al., Phytochemistry* 26:2901-2903 (1987); Ye *et al., Plant J.* 1:175-183 (1991); Yomo and Taylor, *Planta* 112:35-43 (1973).

A microarray-based method for high-throughput monitoring of gene expression may also be utilized to measure Expression Response. This 'chip'-based approach involves microarrays of nucleic acid molecules as gene-specific hybridization targets to quantitatively measure expression of the corresponding mRNA. Hybridization to a

microarray can be used to efficiently analyze the presence and/or amount of a number of nucleotide sequences simultaneously.

Several microarray methods have been described. One method compares the sequences to be analyzed by hybridization to a set of oligonucleotides representing all possible subsequences. A second method hybridizes the sample to an array of oligonucleotide or cDNA molecules. An array consisting of oligonucleotides complementary to subsequences of a target sequence can be used to determine the identity of a target sequence, measure its amount, and detect single nucleotide differences between the target and a reference sequence. Nucleic acid molecule microarrays may also be screened with protein molecules or fragments thereof to determine nucleic acid molecules that specifically bind protein molecules or fragments thereof.

The microarray approach may be used with polypeptide targets (U.S. Patent Nos. 5,445,934; 5,143,854; 5,079,600; and 4,923,901). Essentially, polypeptides are synthesized on a substrate (microarray) and these polypeptides can be screened with either protein molecules or fragments thereof or nucleic acid molecules in order to screen for either protein molecules or fragments thereof or nucleic acid molecules that specifically bind the target polypeptides.

In a preferred embodiment of the present invention microarrays may be prepared that comprise nucleic acid molecules where preferably at least about 10%, preferably at least about 25%, more preferably at least about 50% and even more preferably at least about 75%, 80%, 85%, 90% or 95% of the nucleic acid molecules located on that array are selected from the group of nucleic acid molecules that hybridize under low, moderate or high stringency conditions to one or more nucleic acid molecules having a nucleic acid sequence selected from the group of SEQ ID NO: 1 through 3, 5 through 47, and complements thereof.

In another preferred embodiment of the present invention microarrays may be prepared that comprise nucleic acid molecules where preferably at least about 10%,

preferably at least about 25%, more preferably at least about 50% and even more preferably at least about 75%, 80%, 85%, 90% or 95% of the nucleic acid molecules located on that array are selected from the group of nucleic acid molecules having a nucleic acid sequence selected from the group of SEQ ID NO: 1 through 3, 5 through 47, complements thereof, and fragments of these sequences.

In a preferred embodiment of the present invention microarrays may be prepared that comprise nucleic acid molecules where such nucleic acid molecules encode at least one, preferably at least two, more preferably at least three, even more preferably at least four, five or six proteins or fragments thereof selected from the group consisting of *gcpE, ygbB, ygbP, ychB, dxs* and *dxr*.

The present invention includes and provides a method for determining a level or pattern of a protein in a plant cell or plant tissue comprising (A) incubating under conditions permitting nucleic acid hybridization: (i) a marker nucleic acid molecule having a nucleic acid sequence that hybridizes to a sequence selected from the group consisting of SEQ ID NOs: 1 through 3, 5 through 47, and complements thereof; and (ii) a complementary nucleic acid molecule obtained from the plant cell or plant tissue, wherein nucleic acid hybridization between the marker nucleic acid molecule and the complementary nucleic acid molecule permits the detection of an mRNA for the protein; (B) permitting hybridization between the marker nucleic acid molecule; and (C) detecting the level or pattern of the complementary nucleic acid, wherein the detection of the complementary nucleic acid is predictive of the level or pattern of the protein in the plant.

The present invention also includes and provides a method for determining a level or pattern of a protein in a plant cell or plant tissue comprising (A) assaying the concentration of the protein in a first sample obtained from the plant cell or plant tissue; (B) assaying the concentration of the protein in a second sample obtained from a reference plant cell or a reference plant tissue with a known level or pattern of the

protein; and (C) comparing the assayed concentration of the protein in the first sample to the assayed concentration of the protein in the second sample.

I. Screening Uses

The present invention provides methods and agents that can be used to screen for and isolate genes associated with the MEP pathway. Because the MEP pathway is an essential pathway, disruption of any essential gene in the MEP pathway will result in the death of the cell or organism. While not being limited to any particular biological process, the present invention provides a method and the agents associated with such a method where mutations that result in loss of function of a MEP pathway gene do not result in cell or organism death by providing a second pathway capable of synthesizing IPP and DMAPP. The present invention provides cells and organisms having a second pathway capable of synthesizing IPP and DMAPP.

In a preferred aspect, a cell or organism comprising: (a) a first DNA sequence encoding an enzyme having catalytic activity of mevalonate kinase; (b) a second DNA sequence encoding an enzyme having catalytic activity of 5-phosphomevalonate kinase; (c) a third DNA sequence encoding an enzyme having catalytic activity of 5-diphosphomevalonate-decarboxylase; and (d) a fourth DNA sequence encoding an enzyme having catalytic activity of isopentenyl diphosphate isomerase; wherein at least two of said first, second, third, or fourth DNA sequences have a foreign DNA sequence.

In a preferred aspect, the second pathway capable of synthesizing IPP and DMAPP has at least one, more preferably at least two, even more preferably at least three or four enzymes selected from the group consisting of: mevalonate kinase, 5-phosphomevalonate kinase, 5-diphosphomevalonate decarboxylase and isopentenyl diphosphate isomerase. In a more preferred embodiment, at least two, even more preferably at least three or four of the enzymes selected from the group consisting of: mevalonate kinase, 5-phosphomevalonate kinase, 5-diphosphomevalonate decarboxylase

and isopentenyl diphosphate isomerase are encoded by a foreign DNA sequence. Any foreign DNA encoding such enzymes may be utilized such as human 5-phosphomevalonate kinase (Genbank Accession No. HO9914).

Any cell or organism that possesses the MEP pathway may be used in this aspect of the invention. By providing a second pathway capable of synthesizing IPP and DMAPP, such cells can be utilized in methods to examine the function of a gene, determine whether a gene is associated with the MEP pathway, and identify a gene associated with the MEP pathway.

The present invention includes and provides a cell comprising: (a) a first DNA sequence encoding an enzyme having catalytic activity of mevalonate kinase; (b) a second DNA sequence encoding an enzyme having catalytic activity of 5-phosphomevalonate kinase; (c) a third DNA sequence encoding an enzyme having catalytic activity of 5-diphosphomevalonate-decarboxylase and (d) a fourth DNA sequence encoding an enzyme having catalytic activity of isopentenyl diphosphate isomerase; wherein at least two of the first, second, third or fourth DNA sequence have a foreign DNA sequence.

The present invention includes and provides a method for examining the function of a gene associated with the MEP pathway, comprising: (a) rendering inoperative the gene in a first cell capable of converting mevalonic acid to isopentenyl diphosphate and dimethylallyl diphosphate; (b) rendering inoperative the gene in a second cell incapable of converting mevalonic acid to isopentenyl diphosphate and dimethylallyl diphosphate; and (c) determining the viability of the first cell and the second cell.

The present invention includes and provides a method for determining whether a gene is associated with the MEP pathway, comprising: (a) rendering inoperative the gene in a first cell capable of converting mevalonic acid to isopentenyl diphosphate and dimethylallyl diphosphate; (b) rendering inoperative the gene in a second cell incapable

of converting mevalonic acid to isopentenyl diphosphate and dimethylallyl diphosphate; and (c) determining the viability of the first cell and the second cell.

The present invention includes and provides a method for identifying a gene associated with the MEP pathway, comprising: (a) rendering inoperative the gene in a first cell capable of converting mevalonic acid to isopentenyl diphosphate and dimethylallyl diphosphate; (b) rendering inoperative the gene in a second cell incapable of converting mevalonic acid to isopentenyl diphosphate and dimethylallyl diphosphate; and (c) determining the viability of the first cell and the second cell.

Application of the teachings of the present invention to a specific problem or environment is within the capabilities of one having ordinary skill in the art in light of the teachings contained herein. Examples of the products and processes of the present invention appear in the following examples, which are provided by way of illustration, and are not intended to be limiting of the present invention.

EXAMPLE 1

ISOLATION AND MUTAGENESIS OF THE CODING SEQUENCES OF THE MVA$^+$ TRANSCRIPTION UNIT

Yeast Diphosphomevalonate Decarboxylase (yPMD, ORF YNR043w, *ERG19*)

The coding sequence of yPMD is amplified by PCR using genomic DNA using *Saccharomyces cerevisiae* strain FY1679 as template. The reaction mixture of the PCR is prepared in a final volume of 25 µl containing 1 µg of template, 0.5 µM of primers CINCO (SEQ ID NO: 51) and SEIS (SEQ ID NO: 52), 100 µM of each deoxynucleoside triphosphate (dNTPs) and *Pfu* reaction buffer (20 mM of Tris-HCI adjusted to pH 8.8, 2 mM of $MgSO_4$, 10 mM of KCI, 10 mM of $(NH_4)_2SO_4$, 0.1 % of Triton X-100, 100 µg/ml of BSA). The sample is covered with mineral oil, incubated at 96° C for 3 minutes and cooled to 80° C. *Pfu* DNA polymerase (1 unit, Stratagene) is added and the reaction

mixture is incubated for 30 cycles consisting of 1 minute at 94° C and 4 minutes 30 sec at 72° C, followed by a final step of 10 minutes at 72° C. The PCR product (1879 bp) is cloned in the *Sma* I restriction site of plasmid pBluescript SK+.

Nde I and *Eco* RI restriction sites are introduced, respectively, at the 5' and 3' end of the yPMD coding sequence by PCR, using plasmid DNA as template. The reaction mixture of the PCR is prepared in a final volume of 50 l containing 200 ng of template, 1 μM of primers MPD-Nde5' (SEQ ID NO: 53) and MPD-Eco3' (SEQ ID NO: 54), 100 μm of dNTs, *Pfu* reaction buffer and 1.25 units of *Pfu* DNA polymerase. The sample is denatured for 2 minutes at 94°C and incubated for 10 cycles consisting of 1 minute at 94° C, 1 minute at 61° C and 2 minutes 30 sec at 72° C. The PCR product (1207 bp) is cloned in the *Sma* I restriction site of plasmid pBluescript SK+. Sequencing is performed to ensure that no additional mutation had been introduced during amplification.

Human 5-Phosphomevalonate Kinase (hPMK)

A *Hpa* I restriction site is introduced at both ends of the coding sequence of the human 5-phosphomevalonate kinase by PCR, using the cDNA clone ym0505.rl from Soares infant brain 1NIB as template. The clone ym0505.rl (I.M.A.G.E. 46897; GenBank accession number H09914) is obtained from Research Genetics, Inc (Huntsville, Alabama). The reaction mixture of the PCR is prepared in a final volume of 50 μl containing 200 ng of template, 1 μM of primers hPMK1 (SEQ ID NO: 55) and hPMK4 (SEQ ID NO: 56), 100 μM of dNTPs, *Pfu* reaction buffer and 1.25 units of *Pfu* DNA polymerase. The sample is denatured for 2 minutes at 94° C and incubated for 10 cycles consisting of 30 sec at 94°C, 40 sec at 65° C and 1 minute 45 sec at 72° C. The PCR product (601 bp) is cloned in the *Sma* I restriction site of plasmid pBluescript SK+ and sequenced.

Yeast Mevalonate Kinase (yMVK, ORF YMR208w, *ERG12*)

The coding sequence of yMVK is amplified by PCR using genomic DNA from *Saccharomyces cerevisiae* strain FY1679 as template. The reaction mixture of the PCR is prepared in a final volume of 25 µl containing 1 g of template, 0.5 µM of primers UNO (SEQ ID NO: 57) and DOS (SEQ ID NO: 58), 100 µM of dNTPs and *Pfu* reaction buffer. The sample is covered with mineral oil, incubated at 96° C for 3 minutes and cooled to 80° C. One unit of *Pfu* DNA polymerase is added and the reaction mixture is incubated for 30 cycles consisting of 1 minute at 94° C and 4 minutes 30 sec at 72° C, followed by a final step of 10 minutes at 72° C. The PCR product (1744 bp) is cloned in the *Sma* I restriction site of plasmid pBluescript SK+.

A *Hpa* I restriction site is introduced at both ends of the yPMK coding sequence by PCR, using plasmid DNA as template. The reaction mixture of the PCR is prepared in a final volume of 50 µl containing 200 ng of template, 1 µM of primers MK-Hpa5' (SEQ ID NO: 59) and MK-Hpa3' (SEQ ID NO: 60), 100 µM of dNTPs, *Pfu* reaction buffer and 1.25 units of *Pfu* DNA polymerase. The sample is denatured for 2 minutes at 94° C and incubated for 10 cycles consisting of 45 sec at 94° C, 45 sec at 57° C and 2 minutes 50 sec at 72° C. The PCR product (1351 bp) is cloned in the *Sma* I restriction site of plasmid pBluescript SK+ and sequenced.

Isopentenyl Diphosphate Isomerase from *Escherichia coli* (ecIDI)

The coding sequence of the isopentenyl diphosphate isomerase from *E. coli* is amplified by PCR, using genomic DNA from strain W3110 as template. In this PCR, a *Xho* I restriction site is introduced at both ends of the coding sequence. The reaction mixture of the PCR is prepared in a final volume of 50 µl containing 200 ng of template, 0.5 µM of primers idi5X (SEQ ID NO: 61) and idi3X (SEQ ID NO: 62), 100 µM of dNTPs and *Pfu* reaction buffer. The sample is covered with mineral oil, incubated at 96° C for 3 minutes and cooled to 80° C. *Pfu* DNA polymerase (1.5 units) is added and the

reaction mixture is incubated for 5 cycles consisting of 30 sec at 94° C, 40 sec at 55° C and 1 minute 45 sec at 72° C and 25 cycles consisting of 30 sec at 94° C and 2 minutes 15 sec at 72° C. The PCR product (569 bp) is cloned in the *Sma* I restriction site of plasmid pBluescript SK+.

EXAMPLE 2

ASSEMBLY OF THE MVA$^+$ TRANSCRIPTION UNIT

The transcription unit is assembled in a derivative of the expression vector pBAD-GFPuv (Clonetech, Palo Alto, California; GenBank accession number U62637). This is a high copy number plasmid that belongs to the pMB1/ColE1 incompatibility group. The final transcription unit is composed of four ORFs coding for yPMD, hPMK, yMVK and ecIDI. The coding sequences are preceded by ribosomal binding sites that consist of a Shine-Dalgarno sequence followed by an AT-rich translation spacer of eight bases (optimal distance to the ATG start codon; Makrides, *Microbiol. Rev.* 60:512+ (1996)). The whole construct is under control of the P_{BAD} promoter, which can be induced in the presence of L-(+)-arabinose and repressed in the presence of D-(+)-glucose and absence of L-(+)-arabinose. Lobell and Schleif, *Science* 250:528-532 (1990); Guzman *et al., J. Bacteriol.* 177:4121-4130 (1995).

As a preliminary step, the *Nde* I restriction site located between pBR322*ori* and the *ara*C coding region of pBAD-GFPuv (position 4926-4931) is eliminated by site-directed mutagenesis as described (Kunkel *et al.*, *Meth. Enzymology* 154:367-382 , 1987), using the oligonucleotide pBAD-mut1 (SEQ ID NO: 63) as mutagenic primer. The mutation is confirmed by restriction analysis and sequencing. The plasmid obtained is named pAB-M0. The GFP coding sequence of pAB-M0 is substituted by the yPMD coding sequence. This sequence was cloned between *Nde* I and *Eco* RI restriction sites,

taking advantage of the modifications introduced at the ends of the yPMD sequence. The yPMD sequence is the first of the transcription unit.

To clone the other coding sequences, a polylinker is first introduced between *Eco*RI and *Sal* I restriction sites. The polylinker is generated by annealing the oligonucleotides pBAD-Link1 (SEQ ID NO: 64) and pBAD-Link2 (SEQ ID NO: 65). It contains the restriction sites *Pme* I and *Sna* BI, flanked by cohesive ends of *Eco* RI and *Sal* I sites. Sites *Pme* I, *Sna* BI and *Sal* I are preceded by the Shine-Dalgarno consensus sequence "TAAGGAGG". The modified inserts coding for hPMK and yMVK are digested with *Hpa* I and blunt ligated, respectively, into *Pme* I and *Sna* BI restriction sites. The modified insert coding for ecIDI is digested with *Xho* I and ligated into *Sal* I restriction site. Insert orientation is confirmed after every step by PCR and sequencing.

The plasmid containing yPMD, hPMK and yMVK is named pAB-M2. The plasmid containing, in addition, ecIDI is named pAB-m3.

EXAMPLE 3

STABLE INTEGRATION OF THE MVA$^+$ TRANSCRIPTION UNIT INTO THE *E. coli* CHROMOSOME

Transfer of the MVA$^+$ transcription unit to the chromosome from *E. coli* is achieved with a genetic system based in two elements: the *E. coli* strain TE2680 (Elliott, *J. Bacteriol.* 174:245-253, 1992) and a pRS550-derived plasmid (Simons *et al.*, *Gene* 53:85-96, 1987). Strain TE2680 is a *recD* (tet') mutant host that allows efficient recombination of a linear (restriction enzyme-cleaved) DNA with homologous sequences present in the chromosome. The new sequence is incorporated as a single copy and is perpetuated through cell division.

The sequence of interest, the MVA$^+$ transcription unit in this case, can be cloned in pRS550 vector, between a functional kanamycin resistance (Kan^R) gene and a

promoterless version of the *lac* operon. A similar cassette is present in the recipient host (strain TE2680), interrupting the *trp* operon. This strain is auxotrophic for tryptophan. In this case, however, a non-functional kanamycin resistance (Kan^S) gene and the deleted version of the *lac* operon are flanking a functional chloramphenicol resistance (Cam^R) gene. A double crossover affecting the *Kan* gene and the deleted version of the *lac* operon substitutes the sequence of interest for the Cam^R gene in the chromosome. As a consequence of the crossover, the recipient strain, originally Kan^S and Cam^R, becomes Kan^R and Cam^S.

The MVA^+ transcription unit is amplified by PCR using the pAB-M3 plasmid as template and oligonucleotides pBAD-D2 (SEQ ID NO: 66) and pBAD-U3 (SEQ ID NO: 67) as primers. The reaction mixture of the PCR is prepared in a final volume of 50 µl containing 200 ng of template, 1 µM of primers, 200 µM of dNTPs, *Pfu* reaction buffer and 1.75 units of *Pfu* DNA polymerase. The sample is denatured for 2 minutes at 94° C and incubated for 10 cycles consisting of 40 sec at 94° C, 50 sec at 59° C and 8 minutes 15 sec at 72° C. The amplified sequence (4126 bp) contains the complete promoter, including the regulatory sequences that respond to arabinose and glucose, and the four ORFs that allow conversion of MVA to IPP and DMAPP, but lacks the transcription termination signals that are originally present in the expression cassette.

A polylinker is introduced in the vector pRS550, to allow cloning of the PCR product containing the MVA^+ transcription unit. The polylinker is generated by annealing the oligonucleotides pRS-L1 (SEQ ID NO: 68) and pRS-L2 (SEQ ID NO: 69). It contains the restriction sites *Pme* I, *Sma* I/*Srf* I and *Not* I, flanked by cohesive ends of *Bam* HI and *Eco* RI sites. Plasmid pRS2110 is generated by cloning the polylinker between *Bam* HI and *Eco* RI restriction sites of vector pRS550. The MVA^+ transcription is cloned in the *Pme* I restriction site of vector pRS2110, with the same orientation than the promoterless *lac* operon, thus restoring transcription of the *lac* operon. The plasmid obtained is named pRS-MVA^+.

Plasmid pRS-MVA$^+$ are digested with *Sal* I and *Sca* I restriction enzymes. This digestion rendered a 3196 bp fragment containing the ampicillin resistance gene and a 13406 bp fragment containing the Kan gene, the MVA$^+$ transcription unit and the deleted version of the *lac* operon. Strain EcAB3-1 is obtained by transformation of strain TE2680 with the linear plasmid DNA. The presence of the MVA$^+$ transcription unit in the chromosome of this strain is confirmed by PCR. The activity of this transcription unit is confirmed by the appearance of blue colonies in plates containing 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (Xgal). Strain EcAB3-1 is resistant to kanamycin (25 μg/ml) and tetracycline (6 μg/ml) and sensitive to chloramphenicol (17 μg/ml) and ampicillin (50 μg/ml). The MVA$^+$ transcription unit is transduced to *E. coli* strain MG1655 using phage P1. The strain obtained is named EcAB4-1.

EXAMPLE 4

IDENTIFICATION AND FEATURES OF THE *gcpE* GENE FROM *E. coli* AND A PUTATIVE HOMOLOG FROM *Arabidopsis thaliana*

To identify genes potentially involved in the MEP pathway, a bioinformatic approach is adopted. Because bacterial genes with related functions are often organized in operons, uncharacterized open reading frames (ORFs) that are beside known genes of the MEP pathway are examined. An ORF of 1195 bp with unknown function is found just upstream of a DXS coding sequence of *Streptomyces coelicolor* (cosmid 6A5, Accession Number AL049485). This ORF is homologous to an essential gene of *Escherichia coli* named *gcpE* (Baker *et al.*, *FEMS Microbiol. Lett.* 94:175-180, 1992 (accession number X64451)). An homolog of this gene, named *aarC*, is identified in *Providencia stuartii* and described as an essential gene involved in density-dependent regulation of the 2'-*N*-acetyltransferase (Rather *et al.*, *J. Bacteriol.* 179:2267-2273, 1997). However, no precise function was assigned to the *aarC* gene.

The *gcpE* gene is broadly distributed in evolution. The occurrence of this gene in completely sequenced genomes strictly correlates with the occurrence of the gene encoding 1-deoxy-D-xylulose 5-phosphate reductoisomerase (*dxr*), which catalyses the first committed step of the MEP pathway. Fourteen out of 26 sequenced genomes contain both *dxr* and *gcpE*. Twelve of these sequenced genomes do not contain *dxr* nor *gcpE*. *The gcpE* gene is also highly conserved in plants. *GcpE* homologs are found as EST entries in *Arabidopsis thaliana* (gb T46582, SEQ ID NO: 5), *Glycine max* (gb AW152929, SEQ ID NO: 6), *Lycopersicon esculentum* (gb AW040413, SEQ ID NO: 7), *Mesembryanthemum crystallinum* (gb AI822799, SEQ ID NO: 8), *Oryza sativa* (gb AA753160, SEQ ID NO: 9), *Zea mays* (gb AW126434, SEQ ID NO: 10), *Pinus taeda* (gb AW042702, SEQ ID NO: 11) and *Physcomitrella patens* (gb AW497432, SEQ ID NO: 12).

A cDNA clone from *Arabidopsis* coding for a *gcpE* homolog (EST clone 135H1T7, accession number T46582) is obtained from the *Arabidopsis* Biological Resource Center (ABRC). This clone encodes a full length protein. The cDNA contains an ORF of 2223 bp that encodes a protein of 740 amino acid residues (SEQ ID NO: 1). The Arabidopsis *gcpE* gene corresponding to this cDNA is located in chromosome V (genomic P1 clone MUP24, accession number AB005246). This gene contains 20 exons that extend along 4 kb of genomic sequence.

Alignment of the *E. coli* and *Arabidopsis gcpE* proteins shows high similarity but also striking differences. The first 75 amino acid residues of the *Arabidopsis* sequence constitute a region that is not present in the bacterial counterpart. A transit peptide for plastids is predicted at this region with the ChloroP V1.0 program accessible at the web site www.cbs.dtu.dk/services/ChloroP/ (Score 0.53295). According to this program, the processing site of the transit peptide would be located between Arg38 and Ser39 (CS-score 2.392). *In vivo* import experiments to chloroplasts demonstrated that the N-terminal region of the *Arabidopsis* protein is a functional transit peptide for plastids.

The putative mature *gcpE* protein from *Arabidopsis* is significantly larger than the *E. coli* counterpart (78 versus 41 kDa). Although the two proteins align and show high similarity at the N- and C-terminal regions, the *Arabidopsis* isoform possesses several additional amino acid sequences between these two regions, particularly a domain of 268 amino acid residues (30 kDa) which is only present in the *Arabidopsis* protein (SEQ ID NO: 1).

EXAMPLE 5

DELETION OF THE *gcpE* CODING SEQUENCE IN THE *E. coli* GENOME

To confirm whether *gcpE* from *E. coli* is indeed involved in the MEP pathway, *gcpE* is deleted in strain EcAB3-1. As mentioned above, mutants of the MEP pathway can be rescued in this strain, in the presence of MVA. Deletion of the *gcpE* gene is accomplished by homologous recombination using construct GC5CAT3 as the donor cassette. In this construct, the *CAT* gene is surrounded by the *gcpE* flanking regions. Substitution of the *CAT* gene for the *gcpE* coding sequence in the genome can be selected by chloramphenicol resistance.

Four PCR reactions are necessary to prepare the GC5CAT3 construct. First, a genomic region of 3231 bp, encompassing the *gcpE* ORF (1116 bp), together with flanking regions, is amplified by PCR, using genomic DNA from strain MC4100 as template. The reaction mixture of the PCR is prepared in a final volume of 50 1 containing 250 ng of template, 0.4 M of primers 1PE (SEQ ID NO: 70) and 4PE (SEQ ID NO: 73), 200 M of dNTPs, 1 mM of $MgSO_4$, *Pfx* reaction buffer and 1.25 units of PLATINUM *Pfx* DNA polymerase (Life Technologies Inc., Rockville, Maryland). The sample is denatured for 2 minutes at 94 C and incubated for 30 cycles consisting of 40 seconds at 94 C, 50 seconds at 67 C and 3 minutes 30 seconds at 68 C.

The regions flanking the *gcpE* coding sequence are amplified by PCR using the PCR product of primers 1PE and 4PE as template. Primers 1PE (SEQ ID NO: 70) and 22PE (SEQ ID NO: 71) are used to amplify the 5' flanking region. In this PCR, primer 22PE generates a *Sma* I restriction site. Primers 3PE (SEQ ID NO: 72) and 4PE (SEQ ID NO: 73) are used to amplify the 3' flanking region. In this PCR, primer 3 PE generates a *Pme* I restriction site. The reaction mixtures of these PCRs are prepared in final volumes of 50 l containing 150 ng of template, 4 M of primers, 200 M of dNTPs, *Pfx* reaction buffer and 1.25 units of PLATINUM *Pfx* DNA polymerase. The samples are denatured for 2 minutes at 94 C and incubated for 10 cycles consisting of 40 seconds at 94 C and 2 minutes at 68 C. The PCR product corresponding to the 3' flanking region (1061 bp) is cloned in the *Sma* I restriction site of plasmid pBluescript SK+. The plasmid obtained is named GC3. Subsequently, the PCR product corresponding to the 5' flanking region (1102 bp) is cloned in the *Pme* I restriction site of plasmid GC3. The relative orientation of the 3' and 5' flanking regions is the same than that in the *E. coli* genome. The plasmid with the two *gcpE* flanking regions is named GC53.

The *CAT* gene is amplified by PCR using the plasmid pCAT19 (Fuqua, 1992) as template and oligonucleotide CAT1 (SEQ ID NO: 74) and CAT4 (SEQ ID NO: 75) as primers. The reaction mixture of the PCR is prepared in a final volume of 50 l containing 100 ng of template, 1 M of primers, 100 M of dNTPs, *Pfx* reaction buffer and 1.25 units of PLATINUM *Pfx* DNA polymerase. The sample is denatured for 2 minutes at 94 C and incubated for 20 cycles consisting of 40 seconds at 94 C, 50 seconds at 53 C and 1 minute at 68 C. The PCR product (960 bp) is cloned in the *Sma* I restriction site of plasmid GC53. The construct obtained is named GC5CAT3. In this construct, the *CAT* gene has the same orientation than the *gcpE* gene previously deleted.

Plasmid containing GC5CAT3 construct is digested with *Hind*III, *Xba* I and *Xho* I restriction enzymes to release the recombination cassette. This cassette is amplified by PCR using oligonucleotides 1PE (SEQ ID NO: 70) and 4PE (SEQ ID NO: 73) as primers.

The PCR product is used to transform electrocompetent cells of strain EcAB3-1. These cells are plated on 2xTY medium containing 1.5 % agar (w/v), 17 g/ml chloramphenicol, 6 g/ml tetracycline, 25 g/ml kanamycin, 0.2 % (w/v) L-(+)-arabinose and 1 mM MVA.

The presence of the *CAT* gene in place of the *gcpE* coding sequence in the genome of transformants is confirmed by PCR using oligonucleotides 0PE and 5PE as primers. The identity of the PCR product is verified by restriction analysis. Oligonucleotides 0PE (SEQ ID NO: 76) and 5PE (SEQ ID NO: 77) are complementary to genomic sequences located outside of the region included in the recombination construct. Analysis of transformants confirms both the absence of the original *gcpE* gene and the presence of the *CAT* gene. The novel strain is named EcAB3-3.

Strain EcAB3-3 can grow only in the presence of MVA. A control strain carrying a disruption of *dxs* gene (EcAB3-2) is also auxotrophic for MVA.

EXAMPLE 6

IDENTIFICATION OF GCPE FUNCTION

Example 5 describes the generation of *E. coli* strain with a deletion of the *gcpE* coding sequence (strain EcAB3-3). In addition to the *gcpE* deletion the strain also carries a MVA^+ transcription unit as described in Examples 1, 2 and 3 which makes it auxotrophic for mevalonic acid or mevalonate (MVA). This strain is used to find out which intermediate accumulates due to the disruption of the *gcpE* gene. The gcpE deletion disrupts the MEP pathway blocking the formation of IPP and DMAPP, creating the need for exogenous MVA to synthesize IPP and DMAPP.

A culture of the *E. coli* strain with a disrupted *gcpE* gene is made in the presence of MVA. After growth, the cells are harvested by centrifugation, washed with culture medium containing no MVA and resuspended for 16 hours in a culture medium containing [^{3}H]ME (Methylerythritol). Thin layer chromatography separation of the

water/ethanol (30:70) extract of the cells affords a radioactive band co-eluting with methylerythritol cyclodiphosphate (isopropanol/water/ethyl acetate, 60:30:10, $R_f = 0.56$). Carrier material is obtained for the latter compound from *Corynebacterium ammoniagenes* treated with benzylviologen. Additional data is collected, suggesting that the radioactive compound might correspond to methylerythritol cyclodiphosphate. On HF hydrolysis, it releases free methylerythritol. Like methylerythritol cyclodiphosphate, it is not affected by alkaline phosphatase, which normally cleaves acyclic diphosphates. This compound is not accumulated by the mva+/dxr- *E. coli* strain with an intact *gcpE* gene. In the latter experiment [^{3}H]ME is incorporated into ubiquinone and menaquinone, which are not labeled in the *gcpE* disrupted strain.

Further conformation of function for *gcpE* will require cell-free assays using radiolabeled methylerythritol cyclodiphosphate as described below.

EXAMPLE 7

GCPE ENZYME ASSAYS

Enzymatic preparation of [^{14}C]methylerythritol 2,4-cyclodiphosphate

The substrate methylerythritol cyclodiphosphate cannot be readily chemically synthesized. Attempts to accumulate the tritiated compound from [^{3}H]ME by the mva$^+$/dxr$^-$/gcpE$^-$ mutant described above result in very low yields. Enzymatic synthesis of [^{14}C]methylerythritol cyclodiphosphate is thus required. This can be achieved using all the known enzymes of the MEP pathway, viz., *dxs*, *dxr*, *ygbP*, *ychB*, and *ygbB*.

Enzymatic syntheses of [^{14}C]-deoxy-D-xylulose-5-phosphate (DXP) and MEP from [^{14}C]pyruvate isotopomers and D-glyceraldehyde-3-phosphate (GAP) are performed using *E. coli* strains overexpressing *dxs* and *dxr* genes. In order to prepare the subsequent [^{14}C]methylerythritol cyclodiphosphate from the [^{14}C]MEP the following scheme is used.

Three *E. coli* strains are generated with each one overexpressing one of the three remaining genes in the MEP pathway, viz., *ygbP* (pQE31-ygbP, pREP4), *ychB* (pQE30-ychB, pREP4) and *ygbB* (pQE30-ygbB, pREP4). Each strain is grown on LB medium containing ampicillin and kanamycin at 37°C overnight. Each culture (2ml) is used to inoculate the same medium (50 mL), which are then grown for 3 hours until a 0.5 OD (600 nm) is reached, then induced using IPTG (final concentration 0.1 mM) for 4.5 hours. After centrifugation, the cells of each culture are resuspended in 100 mM Tris-HCl (3 mL, pH 8) and disrupted by sonication (3 x 30 s with 1 min cooling) at 0 °C. After centrifugation, the supernatant is stirred for 1 hour at 0°C in the presence of a 50% Ni-NTA slurry (1 mL, Qiagen Inc., Valencia, California).

The lysate-Ni-NTA mixture is loaded onto a column and the flow-through is collected. The column is washed twice with 100 mM Tris-HCl (4 mL, pH8) containing 50mM imidazole. The proteins are eluted with 100 mM Tris-HCl (2 mL, pH 8) containing 200 mM imidazole. Additional 100 mM Tris-HCl (1.5 mL, pH 8) is added to each protein, and the resulting solution is dialyzed against 100 mM Tris-HCl (pH 8) containing 20% glycerol. On a 12% SDS-PAGE gel, the 6xHis-tagged MEP cytidylyl transferase (*ygbP*), CDP-ME kinase (*ychB*) and 2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate synthase (*ygbB*) are separated from other cellular components.

Using these pure proteins, [^{14}C]2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate is prepared in a one-pot procedure. In a typical incubation, [^{14}C]MEP (10 µL, 2.27x10^6 cpm, 15.8 µCi/µmol) is incubated with the purified MEP cytidylyl transferase (100 µL, 0.4 mg/mL), 6xHis-tagged CDP-ME kinase (200 µL, 0.15 mg/mL) and 2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate synthase (200 µL, 0.6 mg/mL) solutions in 100 mM Tris-HCl (1 mL, pH 8) containing 5 mM CTP, 1 mM ATP, 5 mM $MnCl_2$ and 5 mM $MgCl_2$. The incubation is performed at 37°C for 10 hours.

An aliquot (3 µL) is analyzed on a silica gel plate eluted with isopropanol/water/ethyl acetate (6:3:1). Radioactivity is monitored with a

PhosphoImager. A single radioactive compound is detected. It coelutes with unlabeled 2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate. No radioactivity is found comigrating with ME-CDP. An aliquot is incubated in the presence of alkaline phosphatase and no [^{14}C]methyerythritol is detected, indicating that no [^{14}C]MEP remained in the incubation mixture.

GCPE Enzyme Test

When purified His-tagged GCPE is assayed with the [^{14}C] 2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate as prepared above there is no reaction product detected. One reason for lack of activity could be that GCPE needs other proteins to form a complex with diverting 2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate into the two branches of the MEP pathway. Because of the genetic link of *yfgB* and *yfgA* with *gcpE* (all three are on the same operon of the *E. coli* genome), it is possible that these proteins could be part of this hypothetical enzyme complex. Thus, an expression plasmid containing the genomic region covering *yfgB*, *yfgA* and *gcpE* is constructed and stably transformed into *E. coli* creating the strain BL21(DE3)pLys[PET-T7-gcpE-yfgA-yfgB]. This strain and the BL21(DE3)pLys[PET-T7] and BL21(DE3)pLys[PET-T7-yfgA-yfgB] or [MVA^{+},gcpE^{-}PQE30-AT-gcpE] strains are grown and induced with IPTG using standard conditions.

In a typical experiment, the *E. coli* strain BL21(DE3)pLys[PET-T7-gcpE-yfgA-yfgB] is grown at 30°C in LB medium (50 mL) containing chloramphenicol (34 μg/mL) and ampicillin (100 mg/mL) until reaching a 0.65 OD (600 nm). Induction is then performed with IPTG (0.5 mM) for 6 hours. The cells are harvested by centrifugation (7000g, 10 min) resuspended in buffer (4 mL, 50 mM Tris Hcl pH = 8, 1 mM PMSF, 1 mM DTT, 5 mM $MgCl_2$) and broken at 0°C by sonication (2 x 30 s, with 1 min cooling). The cell debris is removed by centrifugation (16000 g, 10 min).

The resulting crude cell-free material (130 µL) is completed with buffer (20 µL) and used for the enzyme assays at 37°C for 7 hours and 20 hours with the [^{14}C]2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate solution (50 µL) obtained as described above. Controls consist in the same mixture, but the enzyme preparation is replaced by buffer. After incubation, an aliquot (9 µl) of each assay is analyzed on a silica plate eluted with isopropanol/water/ethyl acetate (6:3:1). Radioactivity is monitored with a PhosphoImager.

For unknown reasons, only the assay with *E. coli* BL21(DE3)pLys[PET-T7-gcpE-yfgA-yfgB] extract is successful. In all assays performed with enzyme preparations from other strains, the entire radioactivity comigrated with unlabeled 2-C-methyl-D-erythritol 2,4-cyclodiphosphate, indicating that no reaction occurred. The TLC migration profile is the same as that observed for the control without enzyme.

In the case of all assays performed with the cell system prepared from the BL21(DE3)pLys[PET-T7-gcpE-yfgA-yfgB] strain, there is decrease of the substrate concentration and the accumulation of a new compound. According to its TLC behavior (R_f = 0.85, isopropanol/water/ethyl acetate, 60:30:10), this compound corresponds to a non-phosphorylated derivative. Such a dephosphorylation is most likely, as the test is perfomed with a crude cell-free system containing probably phosphatases, and as no phosphatase inhibitor was added to the incubation buffer. Dephosphorylation of the reaction product might favor displacement of the reaction, the full consumption of the substrate and finally accumulation of a single major product.

The same compound is obtained when only $MgCl_2$ was present in the assay, suggesting that the cofactors tested are not necessary. It is possible that the fact the product is dephosphorylated *in situ* helped to its accumulation. The dephosphorylated new compound (R_f = 0.56, $CHCl_3/CH_3OH$, 8:2) is characterized by a R_f between those of methylerythritol (R_f = 0.22) and isopentenol (R_f = 0.56). TLC comparison with unlabeled

synthetic carriers indicates that compounds 1 to 9 (shown in Figure 1) do not correspond to the non-phosphorylated new compound.

To fully characterize the dephosphorylated product, a larger-scale incubation (10X) is performed and the residue is acetylated (pyridine/Ac_2O, 10 ml) overnight. After the removal of the reagents, the residue is resuspended in $CHCl_3$ (12 ml) and the resulting precipitate is removed by filtration. The filtrate is concentrated to dryness (836000 cpm, 1.1g) and purified on a silica column (8g) eluted with hexane/ethyl acetate (3:1) and fractions of 5 ml are collected. An aliquot (4 µl) of each fraction is spotted on TLC plates (hexane/ethyl acetate, 3:1) and the radioactivity monitored by PhosphoImager. The radioactive fractions of same R_f are pooled together.

Three radioactive products can be detected: Fraction A (200 mg) contains the acetate of the dephosphorylated new compound (R_f= 0.4), fraction B (20 mg) contains the 2-*C*-methyl-D-erythritol triacetate (R_f= 0.2), and fraction C (100 mg) contains another new compound (R_f= 0.25) which is not yet identified. Fraction A is further purified on a silica column (9g) eluted first with CH_2Cl_2 in order to remove almost all impurities and then with ethyl acetate in order to recover the radioactive product. As previously described, an aliquot (4 µl) of each 2 ml fraction is checked for radioactivity and the radioactive fractions are pooled together, concentrated to dryness and almost pure acetate of the dephosphorylated new compound (1 mg) is obtained.

This compound is analyzed by ^{1}H-NMR and from the resulting spectrum it is concluded that the acetate of the putative dephosphorylated GCPE product could be diacetate of (*E*)-2-methylbut-2-ene-1,4-diol. The spectrum is compared with a reference synthetic diacetate of (*E*)-2-methylbut-2-ene-1,4-diol synthesized by $LiAlH_4$ reduction of methylfumaric acid as previously described for the reduction of 3-methylfuran-2(5H)-one or citraconic anhydride (Duvold *et al.*, *Tetrahedron Letters* 38: 6181-6184, 1997). All signals of the enzymatic product match the corresponding signals in the synthetic standard. Furthermore the coelution of the enzymatic radioactive product and the

synthetic diacetate of (*E*)-2-methylbut-2-ene-1,4-diol is observed (CH_2Cl_2, Rf= 0.25). Therefore, one product of the incubation is identified as diacetate of (*E*)-2-methylbut-2-ene-1,4-diol (Figure 2). This positive identification suggests that the product of GCPE reaction with 2-C-methyl-D-erythritol 2,4-cyclodiphosphate is (*E*)-1-(4-hydroxy-3-methylbut-2-enyl) diphosphate (Figure 3).

EXAMPLE 8

CHARACTERIZATION OF *ARABIDOPSIS* GCPE

Upon identification of the *Escherichia coli gcpE* gene as involved in the trunk line of the MEP pathway for isoprenoid biosynthesis, the available databases are searched for plant homologs. As described in Example 4, clone 135H1 (Genbank accession number T46582) is identified as containing an *Arabidopsis thaliana* cDNA encoding a protein with homology to the product of the bacterial *gcpE* gene. As shown in Figure 4, however, the putative *Arabidopsis* GCPE protein (SEQ ID NO: 79), contains several domains that are absent from the *E. coli* protein (SEQ ID NO: 78). Identical residues are in black boxes and conservative changes in grey boxes. Gaps are indicated with dots. The predicted cleavage site for the plastidial targeting peptide (according to the ChloroP program; genome.cbs.dtu.dk/services/chlorop) is indicated with an arrow (*see* Figure 4).

To determine whether the *Arabidopsis* protein encoded by clone 135H1 is indeed a GCPE protein, a complementation assay is carried out using the *E. coli* strain EcAB3-3. In this strain, which is engineered to synthesize IPP and DMAPP from mevalonic acid (MVA), the chromosomal *gcpE* gene is disrupted by insertion of the *CAT* marker conferring chloramphenicol resistance. Because the disruption of *gcpE* is lethal, mutant EcAB3-3 cells require MVA for growth (*see* Example 5).

For the complementation assay, plasmid pQE-AGH is created by subcloning a *Bgl*II-*Sph*I fragment (coding sequence SEQ ID NO: 80 and deduced amino acid sequence

SEQ ID NO: 81) from clone 135H1 into the *Bam*HI-*Sph*I sites of the pQE30 expression vector (coding sequence SEQ ID NO: 82 and deduced amino acid sequence SEQ ID NO: 83) (Qiagen) (Figure 5). The resulting construct encodes a His-tagged protein (coding sequence SEQ ID NO: 84 and deduced amino acid sequence SEQ ID NO: 85) lacking the N-terminal sequence predicted to be a plastidial targeting peptide with the ChloroP program (Figure 5). Expression from plasmid pQE-AGH is under the control of the IPTG-inducible *T5* promoter. Figure 5 depicts the coding sequences in uppercase, and the deduced amino acid sequences are shown below the respective coding sequences. The predicted cleavage site for the plastidial targeting peptide is indicated with an arrow.

EcAB3-3 cells are transformed with plasmid pQE-AGH and plated on LB plates containing 100 mg/l kanamycin (to select for the MVA operon), 34 mg/l chloramphenicol (to select for the *gcpE* gene disruption), 100 mg/l ampicillin (to select for transformants containing pQE-AGH), 0.04 % arabinose (to induce expression of the MVA operon genes), and 0.5 mM MVA (to be used for IPP and DMAPP biosynthesis). The resulting strain, EcAB3-3(pQE-AGH), is able to grow in absence of MVA at 30°C and 37°C, confirming that MVA auxotrophy can be overcome by the presence of plasmid pQE-AGH. These results demonstrate that the cloned *Arabidopsis* cDNA encodes a protein with the same activity as the *E. coli* GCPE protein.

In order to study whether the truncated *Arabidopsis* GCPE protein cloned in plasmid pQE-AGH is active in converting ME-cPP to the next intermediate of the MEP pathway, the protein is expressed at high levels in *E. coli*. Strains XL1Blue or M15 (Qiagen Inc., Valencia, California) are used for expression under several experimental conditions: growth at 23°C, 30°C, or 37°C and induction with 1 or 0.4 mM IPTG, with unsuccessful results. When strain EcAB3-3(pQE-AGH) is used, however, expression of the cloned protein is detected.

An overnight culture of EcAB3-3(pQE-AGH) cells grown in LB medium supplemented with kanamycin, chloramphenicol, ampicillin, arabinose and with or

without MVA at the concentrations described above is diluted 1:50 in fresh medium and incubated at 37°C until reaching an OD_{600} of ca. 0.3. Although cells grew better when MVA is added to the medium, the presence of plasmid pQE-AGH is sufficient to allow growth in the absence of any exogenous source for isoprenoid synthesis. Expression of the truncated *Arabidopsis* GCPE protein is induced by adding IPTG to a final concentration of 0.4 mM.

After incubation at 30°C for 4 hours, cells are collected by centrifugation and resuspended in a 1/50 volume of homogeneization buffer (Tris-HCl 20 mM pH 8.0, 1 mM β-mercatoethanol, 1mg/ml lysozime, 80 mg/l PMSF, and 1 tablet/20 ml of Complete Mini, EDTA-free Protease Inhibitor Cocktail Tablets (Roche Molecular Biosystems, Indianapolis, Indiana)). Following incubation at room temperature for 20 minutes, cells are sonicated 5 times for 30 seconds at 30W. The insoluble fraction is pelleted by centrifugation at 5000x*g* for 30 minutes and the supernatant (soluble fraction) is collected. Electrophoresis on SDS-PAGE of an aliquot of this soluble fraction shows that a protein of the expected size (ca. 78 kD) is expressed in cells grown with or without MVA.

Purification of the His-tagged protein from the soluble extract is carried out using HiTrap columns (Pharmacia, Uppsala, Sweden). Flux through the column is kept constant at 2.5 ml/min during all the steps. After applying the sample to a column and washing unbound proteins with 20 ml of washing buffer (20 mM Tris-HCl pH 8.0, 10 mM imidazole, 500 mM NaCl), elution is performed with 50 ml of a gradient solution containing from 10 mM to 500 mM imidazole and 2.5 ml fractions are collected afterwards. The truncated *Arabidopsis* GCPE protein elutes at 100 mM imidazole and is virtually pure.

EXAMPLE 9

PREPARATION OF PLANT EXPRESSION VECTORS WITH GCPE

Rice, soybean and *E. coli gcpE* genes are chosen for plant expression. An *E. coli* gene (SEQ ID NO: 3) is cleaved by *Nco*I / *Eco*RI restriction digest, gel purified, and ligated into *Nco*I / *Eco*RI-digested and gel purified pMON26541 resulting in the formation of a shuttle vector. These ligations fuse the bacterial *gcpE* gene to CTP1, which is the chloroplast target peptide of the small subunit of the ribulose bisphosphate carboxylase from *Arabidopsis*, and place it under e35S promoter control.

To place the *gcpE* gene under napin promoter control, the shuttle vector is digested with *Eco*RI, ends are filled in using the Klenow fragment, and the gel purified vector is digested with *Bgl* II. The smaller fragment encoding the *gcpE* gene fused to CTP1 is gel purified. pCGN3224 is digested with *Pst*I, ends are filled in with Klenow fragment and subsequently the vector is digested with *Bgl* II and gel purified. The purified vector and the purified CTP1::*gcpE* fusion are then ligated into digested and gel purified pGCN3223.

To transfer the *E. coli gcpE* gene into an *Arabidopsis* binary vector, pGCN3223 is digested with *Hind*III and *Sac* I and the gel purified fragment carrying the e35S promoter fused to CTP1 and *gcpE* is ligated into *Hind*III / *Sac*I-digested and gel purified pMON26543, resulting in a vector containing *gcpE* under e35S promoter control. The pNapin binary expression vector is obtained by ligating the gel purified *Not*I fragment harboring the pNapin::CTP1::*gcpE*::napin 3' expression cassette into *Not*I digested pMON36176.

Seed-specific expression vectors for a rice *gcpE* (SEQ ID NO: 2) and a soybean *gcpE* (SEQ ID NO: 6) sequence are constructed using a pBin19 (Bevan, *Nucleic Acids Research* 12: 8711-8720, 1984) derivative. The plasmid contains the *Vicia faba* seed-specific promoter from the Legumin B4 gene (Bäumlein *et al., Nucleic Acids Research*

14: 2707-2719, 1996), the sequence encoding the transit peptide of the Nicotiana tabacum transketolase (TkTp) (R. Badur, Ph.D. thesis, Georg August University of Göttingen, Germany, 1998) and the transcriptional termination sequence from the octopin synthase gene (Gielen *et al., EMBO J.* 3:835-846, 1984). A rice *gcpE* (SEQ ID NO: 2) sequence is cloned in sense orientation as a *Bam* HI fragment into the *Bam* HI site of the pBin-LePTkTp9 vector, resulting in a recombinant rice *gcpE* expression vector. A recombinant soybean gcpE (SEQ ID NO: 6) expression vector is similarly created.

EXAMPLE 10

TRANSFORMATION OF PLANTS

Agrobacterium transformed with the vectors of Example 9, and with pQE-AGH (which contains the *Arabidopsis gcpE* gene), are prepared as follows. 100µl of an overnight culture is spread on an agar LB plate with antibiotics. The plate is placed upside down in a 30°C chamber overnight. The plates are removed after colonies have grown (24-48 hours). A small scale culture is started by placing 10 ml of liquid LB media in a 50 ml tube. 10µl Kanamycin (50 µg/µL), 10µl Spectinomycin (75-100 µg/µL), and 10µl Chloramphenicol (25 µg/µL) are added. *Agrobacterium* is added from a plate, and the tube is shaken and placed in a 30°C shaker overnight.

Following overnight growth of the 10 ml culture, the culture is removed to a 500 ml flask. 200 ml of liquid LB is placed in a flask, 200µl Kanamycin (50 µg/µL), 200µl Spectinomycin (75-100 µg/µL), and 200µl of Chloramphenicol (25 µg/µL) are added, and the entire 10ml overnight culture is then added. The 500 ml flask is placed in a 30°C shaker and grown overnight. The entire 200 ml culture is placed in a centrifuge tube and centrifuged for 25 minutes at 3,750 rpm and 19°C. After centrifugation, the liquid is poured off and the pellet is resuspended in 25 ml of 5% Sucrose (0.05% Silwet) solution.

900µl of the sucrose solution and 100µl of the 25 ml bacterial culture are placed in a cuvette, and the cuvette is shaken with a covering of parafilm. A blank OD reading is taken with 1 ml of sucrose solution, and then readings of all the bacterial solutions are taken. The OD (at a wavelength of 600) of each culture is recorded. The following calculations are then performed: $C_1V_1 = C_2V_2$; $C_1V_1 = (0.8)(200ml)$; $C_1V_1 = 160$; $V_1 = 160 / C_1$; and $V_1 = X$ ml/10 to determine $OD_{600} = 0.8$ of an *Agrobacterium* culture.

Plants are soaked for at least 30 minutes in water prior to dipping. The bacterial solution is poured into a shallow plastic container, and above ground parts of the plant (bolts, rosettes) are dipped into the solution for 3-5 seconds with gentle agitation. Dipped plants are placed on their side in a diaper lined black tray, and covered by a dome overnight (16-24 hours) to maintain a high humidity. The cover is removed and normal plant growth conditions are resumed for 4 weeks.

Following the transformation and high humidity treatment, plants are maintained at 22°C, 60% RH, and a 16 hour photoperiod for 4 weeks. 5-7 days after transformation, plants are coned. Fertilization with a weak 20-20-20 fertilizer is done weekly. After 4 weeks of growth, plants are placed in the greenhouse and all watering is stopped to encourage plant dry down for seed harvest. Plants are ready for seed harvest after 1-1.5 weeks of dry down. Seeds are harvested by cutting the base of the plant below the cones, holding the plant over a seed sieve and a white piece of paper, running bolts through the cone hole, and collecting clean seeds through sieving.

Seeds are sterilized by connecting a vacuum desiccator hose to a vacuum in a fume hood/flow bench. 100 ml of bleach is placed in a 250 ml beaker, and 3 ml of concentrated HCl is added to the bleach. The beaker is placed in the desiccator, and seeds in seed tubes in a tube holder are placed in the desiccator. A cover is placed on the desiccator, and the vacuum is operated. The desiccator is left overnight but no longer than 16 hours.

Once sterilized, seeds are plated on selection media (prepared by adding 10g (2g/L) Phyta-Gel, 10.75 g (2.15 g/L) MS Basal Salts (M-5524 from Sigma), 50 g (10g/L) sucrose, and 6 ml (1.2 ml/L) Kanamycin solution (950mg/ml), 5ml (1ml/L) Cefotaxime Solution (250 mg/ml), and 5 ml (1 ml/L) Carbenecillin Solution (250 mg/ml) to a total volume of 5 liters at a pH of 5.7). Seed tubes are tapped lightly over a plate in order to distribute the seeds sparsely. The plates are wrapped in parafilm and placed in a 4°C refrigerator for 1-2 days of cold treatment. After this cold treatment the plates are placed in a 28°C chamber for germination.

Selected plantlets are green and have secondary leaves developing. The selected plantlets are moved to soil after secondary leaves have developed. The plantlets are potted in soil and covered with a dome for 5 days to maintain high humidity. The plantlets are moved to a greenhouse after the bottom siliques begin to turn yellow.

Seeds from the selected plantlets are grown in 2.5 inch pots with soil (½ Metro-200; ½ PGX Mix). The soil is mounded and the pot is covered with mesh screen. The screen is fastened to the pot with a rubber band. Seeds are sown and covered with a germination dome. The seedlings are grown in a 12 hour photoperiod in 70% relative humidity at 22°C. Water is supplied every other day as needed and Peter's 20-20-20 fertilizer is applied from below, bi-weekly.

EXAMPLE 11

PRODUCTION OF SEEDS FROM TRANSGENIC PLANTS

Transgenic seed plants from Example 10 representing 20 independent transformation events are grown and seeds harvested to produce T_2 seeds. The T_2 seeds are grown and tested for tocopherol levels. Tocopherol levels are determined by adding 10 to 15 mg of *Arabidopsis* seed into a 2 mL microtube. A mass of 1 g of 0.5mm microbeads (Biospecifics Technologies Corp., Lynbrook, NY) and 500 µl 1% pyrogallol

(Sigma Chem, St. Louis, MO) in ethanol containing 5 μg/mL tocol, are added to the tube. The sample is shaken twice for 45 seconds in a FastPrep (Bio101/Savant) at a speed of 6.5. The extract is filtered (Gelman PTFE acrodisc 0.2 μm, 13 mm syringe filters, Pall Gelman Laboratory Inc, Ann Arbor, MI) into an autosampler tube. HPLC is performed on a Zorbax silica HPLC column, 4.6 mm x 250 mm (5 μm) with a fluorescent detection using a Hewlett Packard HPLC (Agilent Technologies, Palo Alto CA). Sample excitation is performed at 290 nm, and emission is monitored at 336 nm. Tocopherols are separated with a hexane methyl-t-butyl ether gradient using an injection volume of 20 μl, a flow rate of 1.5 ml/min, and a run time of 12 min (40°C). Tocopherol concentration and composition is calculated based on standard curves for α, β, δ, and γ-tocopherol using Chemstation software (Agilent Technologies, Palo Alto CA).

EXAMPLE 12

TRANSGENIC PLANTS WITH GCPE AND OTHER TOCOPHEROL BIOSYNTHESIS GENES

Canola, *Brassica napus* and soybean plants are transformed with a variety of DNA constructs using a particle bombardment approach essentially as set forth in Christou (1996) or using *Agrobacterium* mediated transformation. Two sets of DNA constructs are produced.

The first set of constructs are "single gene constructs" in which the *gcpE* gene is inserted into a plant DNA construct under the control of an arcelin 5, 7S alpha or napin promoter (Kridl *et al., Seed Sci. Res.* 1:209-219, 1991). The products of the *gcpE* gene can be targeted to the plastid by an encoded plastid target peptide such as CTP1 (Keegstra, *Cell, 56(2)*:247-253, 1989; Nawrath, *et al., PNAS 91*:12760-12764, 1994).

A second set of DNA constructs is generated and referred to as the "multiple gene constructs". The multiple gene constructs contain multiple genes each under the control

of a napin promoter and the products of each of the genes are targeted to the plastid by an encoded plastid target peptide, such as a natural plastid target peptide present in the trans gene, or an encoded plastid target peptide such as CTP1.

The multiple gene construct contains the *gcpE* gene and one or more genes for other MEP pathway proteins, including, but not limited to: a *ygbB* gene; a *ygbP* gene; a *ychB* gene; a *yfgA* gene; a *yfgB* gene; a bifunctional prephenate dehydrogenase such as the *E. herbicola* or *E. coli tyrA* gene (Xia *et al.*, *J. Gen. Microbiol.* 138:1309-1316, 1992), a phytylprenyltransferase such as the slr1736 gene (in Cyanobase www.kazusa.or.jp/cyanobase) or the ATPT2 gene (Smith *et al., Plant J. 11: 8*3-92, 1997), a deoxyxylulose synthase such as the *E. coli dxs* gene (Lois *et al.*, *PNAS 95(5)*:2105-2110, 1998), a deoxyxylulose reductoisomerase such as the *dxr* gene (Takahashi *et al. PNAS 95(17)*, 9879-9884, 1998), an *Arabidopsis thaliana* HPPD gene (Norris *et al.*, *Plant Physiol. 117*:1317-1323, 1998), an *Arabidopsis thaliana* GGPPS gene (Bartley and Scolnik, *Plant Physiol. 104*:1469-1470, 1994), a transporter such as the AANT1 gene (Saint Guily, *et al.*, *Plant Physiol. 100(2)*:1069-1071, 1992), a GMT gene (WO 00/32757, WO 00/10380), an MT1 gene, a tocopherol cyclase such as the slr1737 gene (in Cyanobase) or its *Arabidopsis* ortholog, an isopentenyl diphosphate isomerase (IDI) gene, and an antisense construct for homogentisic acid dioxygenase (Sato *et al.*, *J. DNA Res. 7 (1)*:31-63, 2000).

Each construct is transformed into at least one canola, *Brassica napus* and soybean plant. Plants expressing each of these genes are selected to participate in additional crosses. The tocopherol composition and level in each plant is also analyzed using the method set forth in Example 11.

The tocopherol composition and level in each plant generated by the crosses (including all intermediate crosses) is also analyzed using the method set forth in Example 11. Progeny of the transformants from these constructs will be crossed with each other to stack the additional genes to reach the desired level of tocopherol.

Crosses are carried out for each species to generate transgenic plants having one or more of the following combination of introduced genes: *gcpE*, *ygbB*, *ygbP*, *ychB*; *yfgA*; *yfgB*; *tyrA, slr1736, ATPT2, dxs, dxr, GGPPS, HPPD, GMT, AANT1, slr1737, IDI* and an antisense construct for homogentisic acid dioxygenase.

The above description, sequences, drawings and examples are only illustrative of preferred embodiments that achieve the objects, features and advantages of the present invention. It is not intended that the present invention be limited to the illustrative embodiments. Any modification of the present invention which comes within the spirit and scope of the following claims should be considered part of the present invention.